Universidade do Porto
Faculdade de Ciências

Msc. Thesis in Biodiversity, Genetics and Evolution

A trial of karyotypic microdissection as an enrichment pathway for next-generation sequencing

Author:
Tiago Carvalho

Supervisor:
Stuart J. E. Baird, PhD

Co-supervisors:
Rui Faria, PhD
Catarina Pinho, PhD

September 2012

Acknowledgements

I would first like to personally thank Stuart Baird, for accepting to be my supervisor, and granting me the opportunity to work on this project. I truly feel grateful to have worked with you, and thank you for all the support and guidance.

I am also thankful to my co-supervisors, Catarina Pinho and Rui Faria. Thank you both for all the dedication you invested in this thesis, as well as the constant support and discussion throughout its course. This thesis would surely not be the same without your assistance.

It was a great pleasure to work with you, and I am indebted to all three of you for the great amount of time put into this project, and your immense patience to guide me through it.

I am also obliged to thank Martina Pokòrna, Paula Campos, and Tom Gilbert for their contribution to this project. Without you it would not be possible.

I also take this opportunity to thank the latter two, and all my former colleagues in Denmark, for making my short period there worthwhile.

I also acknowledge the Director of the Biodiversity, Genetics, and Biodiversity master programme, Prof. Paulo Alexandrino, and all its lecturers, for making it a real learning experience.

To all my CIBIO friends and former coworkers, I would like to show my appreciation for the friendship, and all the help I received anytime I felt lost during the lab work. Thank you all also for making the tenuous 30 minute daily ride to CIBIO, and the time spent there, really enjoyable.

To my family I would like to thanks for their love and encouragement, and for always being so understanding. Thank you also for always giving me the freedom to learn from my choices.

To all my good friends in Porto, in particular to Diana M., Miguel, Susana, and Guilherme, I thank you for the encouragment words and the good moments. I would equally like to thank my friends from Viseu, Mafalda, Rita, Pedro, Maia, Diana L., Joana P., Ana S., Joana G. and Carla for their friendship of many years, and for always being there when I need.

This thesis is dedicated to all of you.

Resumo

Com os recentes avanços tecnológicos no campo da genómica, graças sobretudo ao desenvolvimento de plataformas de sequenciação paralela massiva, as ciências biológicas encontram-se em plena revolução. Esta tecnologia, que é também designada de sequenciação de próxima geração, apresenta como principal característica a capacidade de execução de múltiplas reações em paralelo não só num curto espaço de tempo, mas também a um preço por base substancialmente mais baixo do que os métodos de sequenciação previamente disponíveis. A sequenciação de genomas completos, que até há uns anos constituía um feito enorme, como ilustrado pelo esforço exigido na sequenciação do primeiro genoma humano, é então hoje em dia bastante mais acessível.

No entanto, embora existam bastantes vantagens na utilização de sequenciação de próxima geração, as restrições não só ao nível do tamanho das sequências obtidas, assim como limitações e erros que são característicos às demais plataformas de sequenciação, constituem algumas das desvantagens. Estas podem levar a que a assemblagem do genoma não se realize de forma correcta, se resultar na colocação das peças com erros nas regiões erradas, ou se as peças em questão não foram produzidas para todo o genoma.

Nesse sentido, a disponibilidade de ferramentas citogenéticas que permitem dissecar individualmente cromossomas, assim como a existência de métodos de enriquecimento que possibilitam a sua amplificação, facultam a investigação para cromossomas individuais. Este tipo de abordagem permite não só reduzir a complexidade computacional mas também o custo associado à sequenciação e manipulação dos dados sequenciados, aumentando simultaneamente a probabilidade de sucesso da assemblagem completa do genoma através da redução do "ruído"que adviria da presença de outros cromossomas. Assim sendo, poder-se-á considerar que este tipo de metodologia é particularmente adequada para estudos de organismos não-modelo, que não possuem um genoma de referência.

Os répteis pertencentes à ordem Squamata constituem um grupo bastante diverso para o qual este tipo de abordagem pode revelar-se particularmente útil. O estudo desta diversidade, que se manifesta pela presença de vários sistemas de determinação sexual (genotípica ou ambiental), sistemas de cromossomas sexuais homogaméticos e heterogaméticos, diferentes tipos de métodos de reprodução (oviparidade, viviparidade, e ovoviviparidade), e até mesmo reprodução por partenogénese, encontra-se actualmente pouco e mal aprofundado. Neste contexto, a abordagem acima referida, ao permitir que a concentração de esforços incida em blocos cromossómicos importantes, tais como os cromossomas sexuais, pode ajudar a esclarecer o seu papel na evolução e como potenciadores de diversidade. Além disso, dada a inexistência de marcadores ligados aos cromossomas sexuais para um determinado subgrupo monofilético dos répteis Squamata - os lacertídeos - a pesquisa e desenvolvimento deste tipo de marcadores moleculares seria fundamental. Os mesmos constituem não só um método fiável para determinar o sexo de um indivíduo com importantes aplicações em ecologia molecular, mas também, e talvez de forma mais relevante, permitem realizar estudos de genómica comparativa ou de genómica das populações.

Como forma de testar a exequibilidade desta abordagem, vários cromossomas sexuais W pertencentes a uma fêmea de *Eremias velox*, uma espécie de lacertídeo, foram microdissecados, enriquecidos, e

sequenciados com a tecnologia da Roche 454. O trabalho descrito nesta tese empreendeu a tarefa de tentar reconstruí-lo *in silico*, e validar esta assemblagem.

Para esse efeito, uma primeira assemblagem foi realizada com o software Newbler. Contudo, a detecção de uma contaminação bacteriana nos dados sequenciados, que levaria à exclusão de mais de metade dos fragmentos sequenciados, motivou o desenvolvimento de uma aplicação informática capaz de incluir todos os passos que constituem a assemblagem - desde o processamento inicial dos fragmentos até à produção final de contigs - maximizando, assim, o número de fragmentos usados. Os objectivos passariam por evitar a exclusão desnecessária de fragmentos, possivelmente relevantes para na assemblagem do cromossoma W de *E. velox*, mas também para a obtenção de uma maior profundidade e amplitude de cobertura do cromossoma, a fim de produzir um maior número de contigs com elevado grau de confiança. Por outro lado, a não exclusão de fragmentos de origem bacteriana, permitiria usar a assemblagem destes como um controlo interno da validade da mesma.

Com o intuito de validar tanto a assemblagem efectuada pelo Newbler, assim como aquela efectuada posteriormente com a aplicação desenvolvida, procedeu-se à comparação *in silico* dos resultados através da análise do seu mapeamento contra o genoma da bactéria apontada como a principal fonte de contaminação. Adicionalmente, para efeitos da validação experimental dos resultados obtidos, alguns contigs produzidos pela aplicação desenvolvida foram selecionados e usados como modelo para desenho de primers com o intuito de amplificar e sequenciar as amostras de ADN pertencentes ao lacertídeo, pelo método de Sanger. Embora os contigs produzidos pelas assemblagens parecessem ser candidatos promissores, o conjunto de testes preliminares com um pequeno conjunto de loci não resultou em qualquer amplificação das amostras de lacertídeo.

Futuramente, o trabalho passará pela procura de novos contigs candidatos mais adequados, bem como por explorar outros métodos de validação que permitam inferir se algum dos passos, assemblagem de contigs ou métodos de validação de laboratório, poderia ser apontado como a causa principal que pudesse explicar a falta de amplificação. Adicionalmente, dada a forte possibilidade de que a ausência de resultados positivos durante a validação laboratorial seja consequência da presença de um número limitado ou absência total de dados de sequenciação referentes ao lacertídeo, o que resultaria no desenho de primers inespecíficos, dever-se-ão realizar novos esforços de sequenciação reforçando-se as medidas para prevenir contaminação.

Palavras-chave: Enriquecimento Genómico Focal, Amplificação da Totalidade do Genoma a partir de Células Únicas, Sequenciação paralela massiva, Assemblagem do Genoma, Cromossomas Sexuais, Microdissecção por Laser, Microcromossomas, Lacertídeos;

Abstract

The relatively recent technological advancements in the genomics field epitomized by the development of high-throughput sequencing methods have come to revolutionize it. Befittingly labeled as next-generation sequencing (NGS), this technology exhibits as its main feature the capability to perform in parallel multiple reactions, not only in a shorter amount of time, but also at a much lower price per base, when compared to older sequencing methods. As a result, sequencing whole genomes, a major feat at the time the first human genome was completed, is today a substantially more approachable task.

While NGS methods have many advantages, drawbacks such as shorter read size, as well as platform related biases and errors are known to be ever-present issues. These undesirable features may thwart a subsequent genome assembly leading to a misconstrued genome picture where pieces may be misplaced and/or missing.

The availability of cytogenetic tools to microdissect single chromosomes and enrichment methods which permit their amplification give us the means to focus on individual chromosomal units. By reducing the costs and computational complexity commonly associated with this kind of trial, as well as improving the odds of assembly by subtracting the "noise" from other chromosomes, this type of approach is particularly well suited for studies involving non-model taxa lacking a genome reference.

Squamate reptiles, a well-diversified group with a broad spectrum of reproductive and sex determination modes and mechanisms, scatteredacross the entire taxa, tipify a case for which this kind of approach should prove itself particularly rewarding. This diversity which manifests itself by the presence of both genotypical and environmental sex determination systems, homogametic and heterogametic sex chromossomal systems, different types of reproduction methods (oviparity, viviparity, and ovoviviparity), and even parthenogenic individuals, is currently largely understudied. In this context, this approach, which grants an opportunity to focus on important chromosomal blocks, such as sex chromosomes, can help clarify their role in evolution and as diversity enhancers. In addition, given the current lack of sexual markers for a particular monophyletic group of squamate reptiles, the lacertids, availability of these molecular markers would be most valuable. Such markers would not only provide a reliable method to sex individuals, but also and more importantly to perform more detailed comparative genomic studies not only between lacertids but also at a broader scale.

As a proof-of-concept experiment, a female *Eremias velox* lacertid had several sexual W chromosomes microdissected, enriched, and sequenced with Roche's 454 technology. The work described on this thesis undertook the task of trying to reconstruct it *in silico* and validate this assembly.

For that purpose, an assembly was primarily performed with the Newbler assembler. The detection of bacterial contamination in the sequence data, which would lead to the exclusion of more than half of the reads, prompted the development of a pipeline, motivated by the will to maximize the amount of data available for the assembly. The goal being to prevent the loss of possibly relevant data, but also to obtain higher depth and breadth of coverage, in order to produce longer and more reliable

contigs. Additionally, by not immediatly excluding the fragments of bacterial origin, their assembly could be used as an internal control of the assembly's validation.

To validate Newbler's and our own pipeline assemblies, the results were first compared computationally, by analysing their mapping to the bacteria genome found to be the main source of contamination. Additionally, contigs assembled by the pipeline were chosen for further lab validation, by testing primers, designed using as template the selected contigs, and verifying if these performed lacertid DNA sample amplication. Although the contigs produced by the assemblies and used as template for primer design were promising candidates, the set of preliminary tests with a small subset of primers failed for to produce any amplification of lizard DNA samples. Future work should then pursue the search for more suitable contigs candidates, as well as explore other validation methods to infer in which side, contig assembly or lab validation methods, lies the cause responsible for the lack of lacertid amplification. Additionally, due to the strong possibility that the lack of a positive lab validation outcome was a consequence of the presence of a very limited amount, or absence, of lacertid sequence data, which would result in the design of primers unspecific to lacertid, new sequencing efforts should be undertaken in the future, employing additional measures to prevent contamination events.

Keywords: Genomic Focal Enrichment, Single Cell Whole-genome Amplification, High-throughput Sequencing, Genome Assembly, Sex chromosomes, Laser Microdissection, Microchromosomes, Lacertids;

Acknowledgment of contribution to the research work

All steps up to and including the chromosomes laser microdissection were undertaken by Martina Pokòrna, at CIBIO, Portugal and CUP, Czech republic. Whole genome amplification was performed by myself and Paula Campos at Natural History Museum of Denmark, Denmark. NGS read analysis and lab validation were performed by me at CIBIO, with valuable feedback from Stuart J.E. Baird, Rui Faria and Catarina Pinho.

Contents

List of Figures

List of Tables

Introduction

Chapter 1

Next–generation sequencing (NGS)

The sheer amount of genetic data produced since the next-generation sequencing technologies saw the light of day half a decade ago (Margulies et al., 2005) is a clear signal that a revolution in the genomics field is currently taking place. Despite the fact that this revolution yields a lot of answers for current biological questions, and further allows scientists to engage in much more complete and far-reaching scientific studies, it is not without its drawbacks. In that sense to make the most of this novel technology it is necessary to understand what changes it can bring to the field of genomics, and what exactly are its strengths and weaknesses so that meaningful results can be achieved. One of the most notable innovations of this technology is the implementation of a method of "clone" library construction directly from DNA molecules, as opposed to the construction of bacterial clone libraries frequently used for large-scale sequencing with the older Sanger sequencing method (Sanger et al., 1977). By choosing to perform clonal amplification it is possible not only to avoid the cloning bias associated with molecular cloning, which often leads to a misrepresentation of certain genomic regions (Bentley, 2008), but also the laborious tasks of construction and handling required to work with the libraries used. In this newly implemented method libraries are instead generally constructed by first fragmenting the DNA molecules (by sonication, nebulization or other shearing method), after which they are end-repaired and polished to create blunt ends. This is followed by the addition of a single A base overhang to the 3' end of the sequence fragment, and finally ligation with designed adaptors of known sequence to its both ends.

In the steps that follow, each of these fragments are denatured in order to create single strands, and later hybridized to library template molecules, present either in beads or in a flow cell surface. Both of these mediums are replete with amplicons, where amplification can take place either by emulsion PCR (Dressman et al., 2003) (used by 454 and SOLiD), after the encapsulation of the beads by lipid vesicles, or bridge PCR (Mitra and Church, 1999) (used by Illumina).[1] In each case, clusters of the same fragment are formed by amplifying it thousands or millions of times in beads or on flow cell surface, respectively[2]. When beads are used these are either loaded into tiny wells or just ligated

[1] These have been the most extensively used methods, more recently new methods have been developed which won't be described.

[2] In this context Helicos a sequence platform which doesn't use amplification in its process can be considered an exception given that it relies on sequencing single molecules skipping the amplification step. (Milos, 2008)

onto a glass slide surface[3]. Sequencing adaptors are then added to the reaction, and respectively, *sequencing-by-synthesis*, or *sequencing-by-ligation* occurs. The former involves the iterative addition of single or multiples nucleotides, depending on the technology, to a growing DNA chain which is the reverse complement of the template sequence, whereas for the latter fluorescently labeled oligonucleotide probes are hybridized to the template. A commonality between both methods is the parallel sequencing of thousands to billions of amplified fragments, or single molecules. In each cycle of nucleotide additions, the hybridization of these with the template is followed by the emmission of a fluorescent signal, which is registered. Depending on the method either each nucleotide or combination of nucleotides addition generates a particular colour emission, or the flow cell is washed with only one type of nucleotide at each cycle and the colour pattern is uniform[4]. Depending on the consensus of the light signal emitted, a quality score is also attributed. The final result is a profusion of sequenced fragments with different characteristics (e.g. size, nucleotidic content bias) depending on the specificities underlying the method used.

Next-generation sequencing truly represents a paradigm shift in the sequencing realm, by effectively superseding the Sanger sequencing speed and cost limits. The cornerstone of next-generation sequencing, is undoubtedly the simultaneous parallel sequencing of millions of DNA fragments, which in the computational context are commonly designated 'reads', a term that goes back to the days when the sequenced DNA fragments were "read" from an autoradiograph of a gel by a human being (Figure A.1 in the appendix) (Flicek and Birney, 2010).

Due to the above mentioned prized features, it is now possible, in a mere question of hours to days at most, to sequence whole-genomes. In addition, it is done at a much lower cost, and in a less labouriously intensive fashion, comparing to the older Sanger capillary method, the field's standard for nearly three decades.

These methods have been reviewed in the literature (Metzker, 2010), and an illustrative figure of the above mentioned NGS methods can be found in the appendix as figure A.5.

1.1 Genome size, coverage and the importance of enrichment

The magnitude of information contained in a single genome is almost unfathomable. In number of nucleotidic bases, it can vary from thousands (Nakabachi et al., 2006), in taxa such as bacteria, to billions (Pellicer et al., 2010), in multicellular organism such as animals and plants

While genomes with up to one million base pairs, such as those of bacteria, can be reasonably sequenced using Sanger's sequencing method, as genome size increases, the more unsuitable the method is for the kind of task at hand. This method inadequacy follows essentially from the high costs, money and time-wise, that result from using it to sequence a whole genome. More so, if there is an expectation to cover such a genome in an even way, but also with satisfying levels of confidence about its true nucleotidic composition. It should be noted however, that this latter factor isn't such

[3] The use of tiny wells as a medium is employed by the 454 sequencing platform, while the latter medium, glass side surface, is used by both Illumina and SOLiD.

[4] Additionally in the case of the 454 sequencing platform the strength of the signal may also reveal how many nucleotides hybridized, but the sensibility is limited since signal saturation occurs.

a prevalent issue with the Sanger method given that in general it is deemed to produce reliable sequencing results.

A major feat at the time of its conclusion, the sequencing of the first human genome, which is approximately 3 billion bp (base pairs), took 13 years to be completed and was initially funded with 3 billion dollars[5]. Today sequencing technologies such as Illumina (Illumina, 2007) can in a matter of hours sequence as much as 1 billion bp in a single run, an amount equivalent in base pairs to one third of the human genome. As a result, by the end of 2010, at least 2.700 human genomes had been sequenced, and predictions suggested that this figure would, by the end of 2011, increase to more than 30.000 (Katsnelson et al., 2010). This ever-increasing trend in the number and variety of genomes sequenced is correlated with a progressively lower cost, as both the price per sequenced megabase and genome continues to decrease (See graphs A.2 and A.3 in the Appendix for a visualization how these parameters have declined with time). Hence there has been a doubling of the amount of genetic information in GenBank (Benson et al., 2004) around every 18 months since its creation in December 1982, with an initial amount of 680.338 bp (Figure A.4 in appendix shows the yearly growth in number of sequences and base pairs).

Notwithstanding the deluge of genomic information that these methods generate, the particular attributes of the reads produced as a result of the sequencing process should be clear in one's mind to apprehend its real value, and its repercussions. For one thing, it is fundamental to keep in mind that, analogously to the older Sanger method, there are inherent biases in the sequencing library preparation, but also on the actual sequencing of the reads, some of which tend to be platform-specific (Whiteford et al., 2005; Huse et al., 2007; Dohm et al., 2008; Harismendy and Frazer, 2009). These can range from a GC content bias, to non-random error occurrence, or low quality inherent to a specific region on the read.

Moreover, the short length nature of the reads produced by NGS, compared to those of Sanger's sequencing, makes it imperative that these reads are oversampled for two reasons. First, in this way the contiguous reads produced in the assembly step, also known as contigs, can be well-covered depth-wise, i.e. with a sufficiently large number of reads overlapping, so that it is possible to ascertain beyond question the original nucleotidic composition. Secondly, to ensure the genome is well covered breadth-wise, since fragments originating from across the genome with a considerable number of different nearby starting positions are required so that reads can overlap along the genome. In this way, it is possible to mitigate the low connectivity effects arising from the short-read size.

Such issues, combined with the redundant nature of certain reads resulting from the genome repetitiveness at small scales, and the randomness associated with the genomic library preparation, which may fail to capture some genome regions, are responsible for the generation of gaps and possible genome misassemblies.

It should also be remarked that sequencing whole-genomes has yet to reach a level where it is economically feasible to routinely sequence all remaining taxa at an individual level. More so, if uniformity across the genome and depth of coverage is sought, which can be particularly problematic if the taxa under study have large and complex genomes.

Indeed, if genome size can present serious hurdles to the sequencing task and assembly, the same can

[5]More information available at http://www.genome.gov/11006943

be said of genome complexity. Due to sequencing platform biases and limitations, complex genomic regions may either go unsequenced, or impose several constraints to its post-sequence assembly. This is where the real bottleneck lies nowadays, as opposed to the genome sequencing step in the past.

Here again, as in the genome size issue, bacteria, with their compact genomes, have it simpler. While these taxa only have a couple of near-identical repeats which exceed 200 bp (Van Belkum et al., 1998), other taxa carrying more complex genomes have different genome repeat content. Human genomes for example, show to different degrees, depending on the activity of LINE (usually ~4kbp) and SINE (between ~500 bp and ~1kbp) transposable elements, as well as the presence of genome duplications and copy number variation, a higher rate of large and more repetitive regions. These regions can amount to as much as 45% of the total human genome length, and in the more general case of reptiles and other mammals, the fraction of genome with these characteristics ranges from 30% to 50% (Hughes and Piontkivska, 2005). If the sequenced reads overlapping these repeat regions, are not long enough to encompass the entire repeat region, they are, due to their redundancy, limited in the amount of information they can provide about the genome. In this type of cases however, special types of reads, namely paired-end and mate-pairs, can be employed to help resolve some of this repeat generated complexity (Medvedev et al., 2009). These can extend over a larger stretch of the genome, and each end of the pair is separated by a known distance. This allows for repeats to be distinguished, even if one of the ends falls within the repeat region, by using the information provided by the opposite end mapping to a unique region. The limitations and implications associated with these repetitive regions in the assembly success, and how these former type of reads may help, will be more fully explored in subsection 1.3.

Currently, coverage can be used to convey three distinct concepts, which need to be properly distinguished in order to avoid any misinterpretation that may result from the arbitrary use of the term.

Two of these types of coverage have already been briefly mentioned above. These are breadth of coverage, which provides a theoretical or empirical measure of how much of a target is covered in width, and depth of coverage, concerning the number of nucleotides contributing to the assembly of a particular contig. A third type of coverage is fold coverage, which concerns the theoretical expectation of the shotgun sequencing outcome, expressed in terms of average nucleotidic base oversampling across the whole genome.

If every type of coverage is identified accordingly:

- B - breadth of coverage
- D - depth of coverage
- F - fold coverage

And the following parameters are given:

- T - target size

- A - assembly size
- L - read length
- n - number of reads

Each type of coverage can be calculated respectively by:

- B = A/T
- D = (nL)/A
- F = (nL)/T

These are used to assess both the success of the sequencing step, and the difficulty associated with it. While the theoretical fold-coverage is more commonly used to estimate the number of gaps in the target coverage in width, by using a mathematical model based on Poisson statistics developed by Lander and Waterman (1988), the other two report to the actual values of the data obtained.

Fold coverage is particularly relevant due to the current limitations of sequencing technologies. Given the current impossibility of sequencing genomes into unique contiguous reads, due to the short size of the reads when compared with genome size, ingenious approaches were suggested to overcome this limitation. One of such approaches, termed whole genome shotgun sequencing (Staden, 1982), is based on the random shearing of genomes in multiple small fragments better tailored for being sequenced. This is then combined with oversampling of the targeted region, to achieve a better coverage both in width and in depth. The assessment of the required oversampling is usually based on the mathematical model developed by Lander and Waterman, which can predict the amount of oversampling (fold-coverage) required to ensure that each base is covered at least once.

Breadth of coverage mainly depends on the presence of repetitive regions (eg. duplications, transposons, and tandem repeats), and polymorphisms (copy number variation, SNPs), the random chance associated with fragment sequencing, and the sequencing-platform related biases and limitations. Each one of these factors may induce what is commonly termed as gaps in coverage. In the absence of a genome reference, this phenomenon results in the impossibility to assemble the regions afflicted. However, even if a genome reference is available, the capacity of the assembly step to detect and distinguish all polymorphic variants may be affected, in the case these happen to fall on regions with zero, or only partial coverage.

Lastly, depth of coverage has, particularly in the context of NGS, a major impact on any genome assembly. It is an inestimable resource for the process of sifting through the possible sequencing errors, and importantly distinguish these from true polymorphisms. Thus, it is the ultimate validation feature that is available to an assembly.

For example, considering a hypothetical single read which is said to have a 1% of error rate. A 10 fold coverage of this same read, supposing all 10 reads match perfectly along part of their entire length, would lower the error rate associated with this particular alignment to 10^{-20}, making it very improbable, although not impossible, that an error is present in the alignment. The low probability associated with this event is then so small that it is reasonable to accept this as a

measure of the reliability of the alignment. From this example it is possible to understand how the confidence associated with an assembly stems in some way from the the depth of coverage available. Similarly, if this was a diploid individual with a 20-fold coverage of the same region, but containing a polymorphism in a particular nucleotidic position, such that sequencing this region would generate ten reads with a given SNP and the other ten with yet another SNP, the relatively high abundance of two sets of reads each differing only in one position would support the presence of a true polymorphism, and not that of an error. It is important however to set a minimum threshold on the length of the alignment. This follows from the fact that the smaller the read length, the more probability there is of random alignments to happen, making it difficult to confidently differentiate not only errors from polymorphisms, but also a true alignment from a spurious one.

However, there is a theoretical limit on the maximum depth coverage required for an assembly. After such threshold is surpassed the assembly will barely see any significant improvements. This occurs mainly in virtue of the limitations imposed both by the size of the reads, as well from the genome complexity, where particularly repeat saturated regions may play a big part if consistently larger than the reads. Theoretical simulations with assumptions of infinite coverage and error-free reads showed that the assembly of a 4 million bp *E. coli* genome (Blattner, 1997), with 20 bp single reads, could only put as much as 10% of bases in contigs of 10 kbp or more (Whiteford et al., 2005). This simple theoretical simulation shows that there is a limit on the coverage benefits, thus demonstrating that both genome complexity, and read length, also may play a significant role in the success of the assembly.

Taking into account the previous arguments, it is easily perceived that despite the great progress made, sequencing technologies still face some arduous challenges which have yet to be effectively tackled. These challenges mainly concern the production of long, unbiased and error-free reads, and their resolution would significantly improve the genome assembly qualities. Currently because of the NGS limitations, and the lack of effort to complement the large scale NGS projects with Sanger sequencing, although the genomes of several species are now available, less than 80% of the sequenced genomes can be considered effectively reliable (Alkan et al., 2011).

Furthermore, and considering that the high costs associated with sequencing still make it a quite impractical and far from trivial step, the aim should be to focus on a particular region of interest by enriching it. Doing so will maximize the amount of sequenced data which is both reliable and target-specific. Thus, this approach truly represents the best value for money, while at the same time facilitates the assembly task, which currently faces the quixotic undertaking of piecing together all the tiny portions originating throughout a genome.

Genomic enrichment, defined as the capture of a target region of interest (Teer et al., 2010; Mertes et al., 2011), is then, for several reasons, a logical choice prior to sequencing. Firstly, it can be used on different genomic scales, meaning that if there is a desire to study a particular region, it can be directly targeted and optionally amplified. This is valid for individual genes, linkage groups, whole chromosomes, or even a mutually shared region across several chromosomes such as one containing microsatellites or other motifs. A selective enrichment contributes then to a reduction of the noise which would be generated from sequencing non-targeted genomic regions. At the same time it also

reduces the work overload usually associated with assigning sequenced fragments to the respective chromosomal units. Lastly, given that the sequencing is circumscribed to a smaller region, not only will the region of interest be better covered in depth, but also breadth-wise. This lends an additional sense of confidence towards the data produced, and its prospective assembly into large contigs.

The above mentioned arguments are especially preponderant for small scientific groups, which often lack the means to produce and/or analyse whole-genomes. Enrichment can be depicted as a suitable "pathway" to partake on the exciting whole genome exploration, and also to pursue small-scale comparative genomic research. This would enable comparisons between individuals from different populations to take place, which is highly prized, but currently not feasible at the whole genome level. But above all, these arguments lay out the preponderant role of enrichment as a source of reliable and target-specific data.

1.2 Current enrichment pathways

Presently, there is a plethora of enrichment pathways available. Each of these make use of slightly different approaches with the same goal of representing in a thorough way, a specific targeted genomic region

The choice of a pathway depends both on the aim of the research, and the cumulative knowledge about the target sequence. In the case the target sequence is known, a myriad of enrichment options exist which mainly involve array and primer design. The designed hybridrization probes will prime the template sequence, and amplification of the target sequence may, or may not, take place.

However, even if no information is available about the nucleotide composition of the targeted sequence, molecular and cytogenetic techniques can be used to circumscribe this nuisance.

Cytogenetic tools provide convenient ways of distinguishing particular chromosomes, or even regions within chromosomes. They do so by exploring the characteristics of the biological units' cytological structure, permiting its clear identification, so that they can be isolated, and further enriched.

To isolate and obtain specific targets, suitable options include FACS (fluorescence activated cell sorting), a special type of Flow Cytometry (FCM) which sorts chromosomes with a laser system according to their size and fluorochrome affinity. However, this approach lacks sensitivity for the isolation of small chromosomes (Zhou and Hu, 2007). FACS can be combined with microdissection in order to isolate specific chromosomes.The chromosomes are then amplified by degenerative oligunucleotide PCR (DOP-PCR) (Telenius et al., 1992), an amplification method that makes use of degenerate primers, in order to create biotin ligated FISH (Fluorescence In Situ Hybridization) probes for chromosome painting (Griffin et al., 1999; Harvey et al., 2002; Henning et al., 2008). This allows the identification of chromosomal-specific regions which can be further microdissected and amplified.

Aside from the cytogenetic enrichment techniques, there are, as previously mentioned, a large array of molecular enrichment techniques available.

One example of a method that doesn't explicitly involve probe design or require extensive knowledge about the sequence being targeted is Restriction Site Associated DNA (RAD). This method aims to reduce the complexity associated with the sequencing of genomic regions. Initially developed to detect polymorphisms by retrieving regions of the genome and using these to create a microarray (Miller et al., 2007), it can now more effectively target whole genomes which may lack a reference genome, or specific genomic regions common to several individuals from a population by combination with NGS technology. This is both useful for approaches that deal with polymorphism discovery (Stapley et al., 2010), and assembly *de novo* (Etter et al., 2011). It uses a restriction enzyme to cut the genome at a desired frequency, proportional to the abundance of the targeted motif in the genome, and normally produces between 10.000 to 100.000 RAD sequences. Furthermore, since the genome is cut by restriction enzyme on specific regions, the fragments produced will all originate from a few relatively space consistent places as opposed to shotgun sequencing which randomly produces sequences from across the genome. These sequences are then adaptor-ligated and can be retrieved for further processing. The recent development of paired-end sequences for this method yields fragments which are particularly useful for genome assembly (Etter et al., 2011).

Other enrichment techniques which require probe design are also available, of which only one, polymerase chain reaction (PCR) (Saiki et al., 1988), can also be used for reference-less target sequences. This technique requires however, some modifications which are described in more detail on section 2.3.

Besides PCR, the other most popular techniques currently in use are respectively molecular inversions probes (MIP) (Hardenbol et al., 2003), and hybrid capture (Lörincz, 1998), either on-array, or in-solution.

Each one of these methods have several different features which need to be weighed up before settling for one approach, and so accordingly, a brief introduction to the most used approaches follows.

To compare each approach, several parameters should be taken into account. These encompass the DNA amount required, the method's specificity and sensitivity, the ability to reproduce the results accurately, and even coverage of the selected region.

Quite possibly one of the most popular methods of amplification is PCR. This is the standard method for amplification used in a normal NGS procedure, although with minor variations depending on the platform. It uses a pair of oligonucleotid DNA strands which connect to opposite ends of the target sequence, and can be as specific as desired, provided that the target flanking regions are unique enough. While PCR is commonly used to target regions not much longer than a few thousands of base pairs, the use of modified protocols can produce fragments exceeding 20kbp (Hogrefe and Borns, 2011). However, aside from the requirement to previously know the target sequence composition to design probes, which can be overcome by using known adaptors to flank the target and designing probes for these, other limitations exist. In particular straight multiplexing issues can happen when multiple primers are used simultaneously. This is due to the unexpected associations which can occur between them, but also due to the competition for resources between the growing fragments (Edwards and Gibbs, 1994). For both these issues there are strategies which can effectively overcome the inherent limitation and allow for multiplexing to take place (Fredriksson et al., 2007; Meuzelaar et al., 2007). In addition, as they grow in length, long amplicons are progressively less reliable copies of the target (Barnes, 1994). Furthermore, due to selective bias towards certain sequences (Warnecke

et al., 1997; Polz and Cavanaugh, 1998; Acinas et al., 2005), a normalization is required in order to acquire a better sample representation, and it is often the case that several optimizations are required for the PCR both to work correctly and require a lower initial DNA input. Given the high reagent-cost, and in the case of DNA sample paucity, optimization can be problematic. Nonetheless, overall PCR is known to have good specificity and sensitivity levels, and given that its reproducibility is high, it can be seen as a dependable method, capable of covering the target beffitingly.

Another method, MIP, uses molecular inversion probes to capture the target sequences. These oligonucleotid DNA strands are composed of a common linker, and a target-specific sequence on both ends. When the target is captured, it undergoes an inversion in configuration, and a circularization occurs, instigated by the action of a ligase enzyme (Nilsson et al., 1994). Then the non-circularized sequences are digested by an exonuclease, thus reducing the background noise, while the remaining are amplified by using primers aimed at the linker sequence. The specificity employed in the target-specific flankers on the probe, partially dictate how successful this approach will be. The characteristics of the probe make this an optimal method for multiplexing, and permit the concomitant amplification of several samples. Additionally, analogously to the PCR method, this approach deals with genomic DNA as input, instead of shotgun created libraries, which translates into a lower DNA input requirement. However, the total achieved sample coverage uniformity is low compared to the other two methods (Mamanova et al., 2010), indicating the need for further optimization. Overall, the features exhibited by this method, show it is an appropriate enrichment method, but the lack of uniformity observed advises one to use it for cases where there is a larger sample number, but low number of targets, maximizing its potential advantages.

Hybrid capture encompasses the other two most popular methods for enrichment. Even though these are based on the same principle, they differ slightly on the method utilized to perform the enrichment. While one is array-based, the other is solution-based. To capture the targeted sequence, the array approach uses a shotgun fragment library, thus requiring a greater initial quantity of DNA, which is hybridized to a microarray slide containing thousands to millions of immobilized oligonucleotid probes. The sequences which don't hybridize with the probes, i.e. non-specific sequences, are then washed away, and the remaining, i.e. the target DNA, are eluted. While the costs associated with the hardware are considered to be high, the process requires less work and is faster than the standard PCR technique (Mamanova et al., 2010). The solution-based approach attempts to alleviate some of the issues present in the array-based method. It employs an higher-ratio of probes over template, effectively reducing the amount of initial DNA sample needed (Gnirke et al., 2009). Furthermore it overcomes the high costs involved on the array-based approach, by not requiring expensive hardware. Thus, in virtue of the particularities of this later approach, it provides an higher specificity and uniformity on the target capture, and can be more easily scalable.

For more comprehensive comparisons between the methods, and detailed reviews see Garber (2008); Summerer et al. (2009); Turner et al. (2009); Mamanova et al. (2010); Teer et al. (2010).

1.3 The issue of assembly

Sequence assembly is a computational challenge that has as final goal the reconstruction of a particular genetic sequence of interest. To accomplish this task fragments are generated by a particular sequencing platform, thereby concomitant with all the hindrances associated with it. In an utopian realm where sequencing technologies would generate error-free sequence fragments with length equal to that of the sequence of interest, there wouldn't be much of a challenge to accomplish in terms of sequence assembly. However, state-of-the-art sequencing technologies are currently limited in both the size and accuracy of the sequencing reads they can generate, resulting ocasionally in impractical sequence assemblies, particularly for genomes presenting some level of complexity.

In order to alleviate the constraints introduced by such limitations sequencing platforms utilize two important approaches. These are "shotgun-sequencing", a technique by which DNA samples are randomly shreaded physically, and sequence oversampling, which is the amplification and sequencing of the DNA sample, so that each genomic region is sequenced multiple of times. The first, "shotgun-sequencing", breaks DNA into smaller sequentiable fragments, with originate from multiple different different genomic origins. It is this latter propriety that improves the chances of constructing long unique sequence paths, obtained from the inferred hierarchical overlaps. The second, sequencing oversampling, provides a way to properly distinguish errors present in some of the reads from polymorphisms, by contributing of the same regions several times, so that further statistical analysis may lead to error identification. The role of the two can be seen more clearly in figure 1.1).

In the typical greedy algorithm[6] each contig will be assembled by comparing each read in pairwise, and merging them if these overlap. This process is repeated until no more reads can be merged, and should optimally lead to a single contig corresponding to the sequenced chromosome, or part thereof.

However, achieving such a high degree of success in a sequence assembly is not as trivial as it may appear. Due to the high complexity it may involve, it is a challenge which can only be solved in both timely and in an effective manner if undertaken *in silico*.

The characteristics intrinsic to the sequenced reads and the assembly's ultimate goal, make it easy to understand why sequence assembly can be seen as a task akin to that of piecing together a large and complex jigsaw puzzle.

A major factor underpinning the effective completion of a jigsaw puzzle is the presence of a box, where an image depicting the puzzle's solution can be found, or absence thereof. This box can present several states of degradation, to the point that parts of the solution may be visually uninterpretable, or even belong to a somewhat similar puzzle. The puzzle's box analogy draws a parallel to the task of sequence assembly where the box stands for genome reference. In general terms, it is this feature which divides sequence assembly into two categories. These are reference guided assembly, more commonly called mapping assembly, and assembly *de novo*. Mapping assembly is preferably performed with the sequenced taxa's genome reference (original box), or otherwise using genomes

[6]Greedy algorithms are probably the simplest approach to the assembly problem. Other algorithms with slight variations exist. For a better overview of the existing algorithms see (Miller et al., 2010)

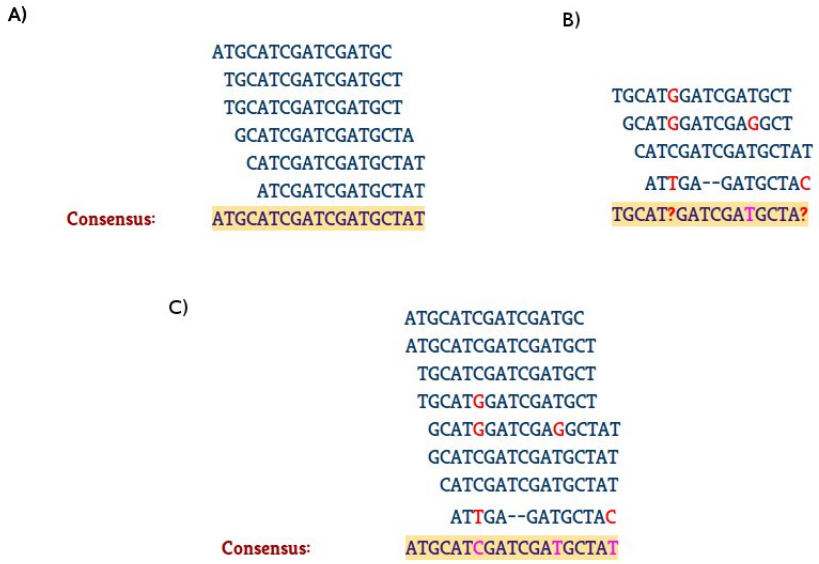


Figure 1.1: **Hierarchical sequence alignment A)** The overlap of reads which share part of their nucleotidic bases gives rise to a consensus contig summarizing their information. Each read when shown not overlap totally with another, effectively extends it, since the outcome is a longer fragment. In this particular example the contig seems to be well covered, mainly at its central portion, and lacking any disagreements between reads, which confers some degree of confidence to its assembly. **B)** In the absence of total agreement between reads, and sufficient coverage of every contig's position, a complete and reliable consensus can be unattainable. In this example, nucleotides in the reads which differ from the consensus in **A**, are marked in red, and dashes represent deletions. In the new consensus, question marks represent positions whose identity couldn't be decidly inferred, and pink characters represent the best guess at the true contig's nucleotide identity, given the presence of more than one option in the reads. The variation present in the reads may be due to errors or sequence polymorphism. Due to the lack of sufficient coverage, a definite guess could not be confidently made for some of the contig's positions (marked with questions marks), while in others faced with the one option for a single position, the most probable nucleotide given its frequency was suggested (marked as pink characters). **C)** While in this example, which includes the reads from **B** in addition to other four reads, it is still not totally clear if all the uncalled positions in **B** were errors or polymorphisms, the level of confidence on the consensus contig generated has increased, and two extra calls were made (7th and last positions of the consensus). Depending on the known read's error rate, and on the importance of distinguishing errors from polymorphisms, if necessary more depth of coverage could be sought to arrive at a more confident consensus.

from taxonomically related or the same taxa (box from similar puzzle). The success of this assembly will vary depending on the completion level and reliability of this reference (how well conserved the box is). The availability of such information grants a way to more aptly piece together the reads and order them more accurately and efficiently by aligning these to a genome reference and seeing where they map[7].

On the other hand, assembly *de novo* relies solely on the information provided by the generated reads, that is without any guiding reference. Often, this may be the only available option for studies focusing on non-model taxa, such as the lacertid dealt with within this thesis, due to the lack of sequenced genomes for comparison. Due to the strict dependence of assembly *de novo* on the reads information, factors such as genome complexity and lack of coverage additionally increase the difficulty associated to this challenge. This stems not only from trying to correctly piecing together the multitude of reads, but also from the extra effort required to blindly validate these connections. Without a genome reference the presence and amount of missing or even mis-assembled contigs may be impossible to ascertain, as these particular instances are usually imperceptible to the researcher's eye just by looking at the contigs.

Still considering the jigsaw puzzle analogy, the assembly process is further complicated by its pieces' characteristics. In particular the reads can be compared to that of a peculiar jigsaw puzzle, where pieces are small, can only be pieced together by a partial overlap between the pieces instead of just locking, a portion of the pieces are missing or damaged, can represent either a normal or a reverse version of the puzzle, some of them were faultily designed and represent two pieces in a single one, and lastly there may be pieces which don't even belong to the puzzle in question. The resolution of this puzzle poses a great number of difficulties, which can be ameliorated by adding an extra set of reads from other boxes containing the same puzzle whence some of pieces might differ slightly, i.e. the result from another sequencing effort for the same or related taxa. However, this comes at a cost since more time will be necessary to compare each piece.

Ultimately unless the puzzle complexity is either low, or there is a good balance between the reads quality and length, there will be mis-assembled pieces. This may results from the pieces' redundancy, i.e. lack of read resolution to decidedly position them in a particular genomic region, the presence of errors, or due to the undesired presence of pieces from other puzzles.Furthermore, as there can be some pieces missing from the initial set, without a reference the puzzle cannot be completed unless new pieces are generated.

This puzzle pieces analogy gives an intuition of the problem related to hypothetical contamination, and conveys a modest idea of why the features particular to each sequencing platform can pose some problems to the assembly. These were described above in the faulty puzzle analogy, and translate into the sequenced fragments read-length, base-call and homopolymer errors, the gaps in the coverage, presence of both forward and reverse read orientation, and chimeric reads, alongside with the massive production of reads.

An additional factor complicating the assembly challenge, and a main cause for the scarcity of finished genomes, is genome complexity. This can be as well represented by yet another jigsaw puzzle analogy, where the complexity can be equated to the blue areas in a puzzle depicting a picture of an almost

[7]The success of this approach relies in the similarity between the sequenced and reference genome at both nucleotidic composition and lay-out. In the presence of differences, misassemblies can be produced unbeknowst to the researcher.

completely blue sky, for which there may not be a reference to follow. In such case, it would be possible to place each blue piece in several positions, without a way to verify if each is correctly positioned.

Equivalently, in the event that this worst-case scenario should happen in a genome, it is not certain that a reasonable sequence assembly would be attainable. Genome complexity, in the assembly context, refers to genomic regions where tandem repeats, duplications, and transposons, among other repetitive elements exist.

Whilst the short size of the reads generated by NGS may not be detrimental to the assembly step, provided they originate from genomic regions with a high degree of uniqueness, the same cannot be said if they derive from complex regions, namely either from the immediate adjacency or inside of repeat filled genomic regions (figure 1.2). When totally contained within the latter regions, reads, when assembled, will most likely collapse into an unique contig. This may happen even if they belong to different genomic regions, provided that both regions present similar repeat patterns. Furthermore the size of such contig may be severely underestimated if the repeat pattern, captured by the reads, is found to be self-repeating over the a particular complex genomic region. Often if this happens the result is a contig whose size may be no longer than the longest read encompassed by the region. Otherwise, in the case that reads originate on the adjacency of similar repeat regions, these may be erroneously merged if they overlap over the shared repeat pattern, which may happen in cases such as genomic duplications.

Ultimately, however, while the amount of complexity in a genome may constrain the assembly, the degree of constraint depends strictly on each reads' size.

To better explore this notion, it is convenient to look at the extreme cases of read length and their informative value. This can be done by simulating a genome, randomly created, and dividing it into all the possible polymer of n nucleotides, or n-mers, starting at length of two [8].

If the set of reads available consists solely of n-mers of size two, it can be easily inferred that no reliable assembly can possibly be produced. In this particular scenario, the overlap between reads, if partial, would result in the extension of a read by at most one nucleotide and, if total, no extension at all. The physical overlap limitation of at most one nucleotide, combined with the random occurrence of nucleotides across the genome, means that no unambiguous contig could ever be produced with absolute confidence. In this case, by chance, most reads will align with high probability, even if they originate from distinct regions in the genome. If the size of the n-mers in this example is progressively increased, the fraction of unique reads, i.e. reads which only appear once in the genome, will naturally increase. Intuitively, this results in a decrease of the amount of possible false overlaps between reads, i.e. occurring by chance, since the probability of two reads sharing most of their nucleotides, if unrelated, decreases as n increases. The progressive reduction of the overall level of reads' ambiguity as their size increases, allows for a more able distinction of true from false overlaps, improving the chances of producing unambiguous true contigs. If the size of the reads is continually increased, ultimately, the read's length will match that of the genome. In this extreme case of the example the fraction of unique reads is 100%, or put in another way, one unique sequence is enough to cover the whole genome.

[8] N-mers containing only one nucleotidic base cannot possibly add any information to a sequence assembly and thus are not considered.

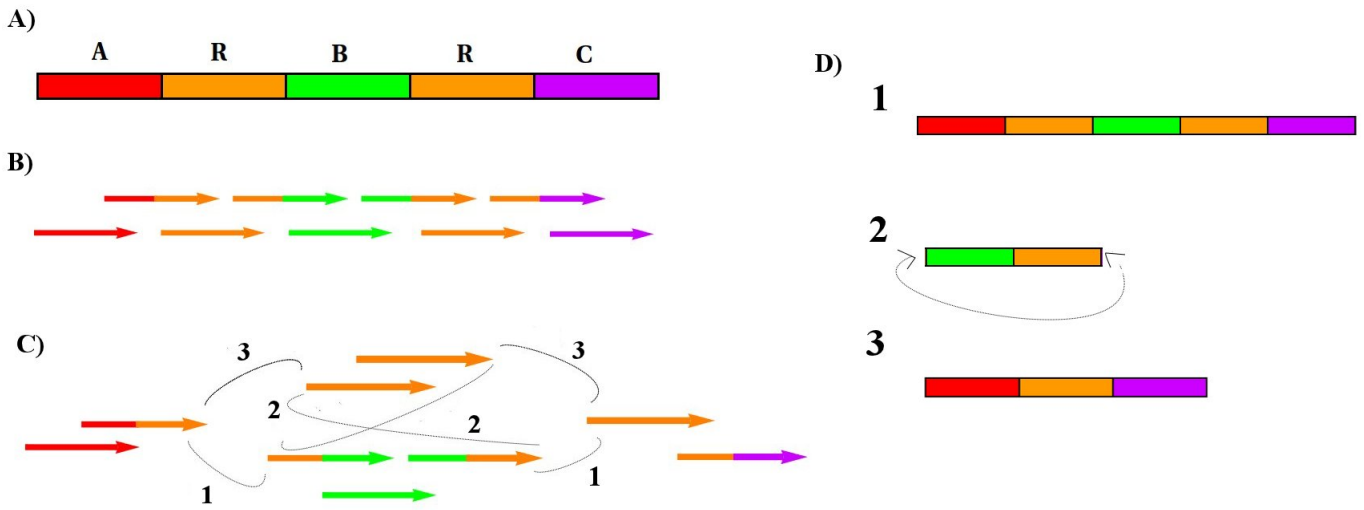


Figure 1.2: **Assembly in presence of repeats. A)** Original sequence with two repeat regions. **B)** Fragments produced by randomly shearing the original sequence. Arrows represent the orientation of the reads (all possess the same orientation). **C)** Three possible assembly paths as indicated by the numbers, in the presence of the sequences in *B*. **D)** The contigs corresponding to the paths in *C*. Only the first case represents a correct assembly of the original sequence. The combination of lack of resolution of some sequences, and the presence of repeats may produce misassembled contigs as those shown here, where the contig #2 is a "circular" contig, links to itself, and #3 missed to include a region inbetween repeats. Figure adapted from A. Policriti.

An important feature of the trend that correlates read length increase and the progressive approachability of their unequivocal assembly, is that it increases exponentially. That is, small increases in read length rapidly translate into verifiable improvements of the assembly step (Schatz et al., 2010).

However, contrary to one of the previous simplifying assumptions, genomes are not entirely random. They are the byproduct of billions of years of evolution, and accordingly have accumulated specific patterns. Such patterns, such as tandem repeats or region duplications, may hardly differ at all, and are repeatedly present throughout the genome. Thus, whilst longer reads greatly improve the chances of unambiguously producing contigs, the exponentiality of the previously mentioned trend, will greatly depend on genome complexity. In normal conditions, the purported trend will not be so pronounced were the genomes to be completely random, and therefore contain every possible nucleotidic combination in equal numbers.

An additional conclusion that can be extracted from the previous thought experiment, is that even with infinite depth genome coverage, if the reads are short, any reasonable assembly will be out of reach. That is, independently of how much the genome is sequenced breadth and depth-wise, with excessively short reads not much can be inferred about the sequence of interest. This is not to say that depth of coverage is deemed as superfluous. In fact, not only is it useful for error-resolution and to confidently connect the reads, but it also confers the ability to identify repetitive regions, which often produce contigs exhibiting inordinate depth of coverage levels in relation to other contigs, thereby allowing them to be more aptly and promptly identified.

While the high-throughtput sequencing technology has been progressing steadily since 2005, with increasingly higher throughtput and longer reads at a signifcantly lower cost, even the longest single reads now produced by Roche's 454 GS FLX Titanium XL+ system [9] lack enough length to resolve some of the longer repeat regions, or part of the genome regions will fail to be sufficiently covered, especially when the genome shows some degree of complexity.

Fortunately, as the NGS technology advances, wet-lab techiques which were initially only available for the Sanger sequencing method are also available for NGS technologies. Of particular interest for assembling complex genomes are the mate-pairs and paired-end reads. These provide auxiliary information to the assembly, by being able to span medium to large genomic regions, allowing the assembly to be better validated.

These reads differ from the single reads in the sense that only the segments' ends, which are separated by a large insert of pre-determined length, are sequenced. Additionally, both show the ability to connect contigs or reads, whose extension was inhibited by either being in the adjacency of long repeat regions, or regions lacking coverage or not sequenced. Although serving the same ultimate purpose, they differ in their generative process, properties, and applicability. Mate-pairs, which are able to encompass larger genomic regions, are reads created by fragmenting the targeted genetic sequence, and circularizing size-selected inserts by linking an internal adaptor. These circularized elements are then randomly sheared, and the fragments containing the adaptor are purified, after which they are sequenced. Even though they are expensive to get, slow to obtain, prone to statistical errors, and incapable of spanning regions longer than 20kb, they are still quite relevant in the actual sequencing context. In particular they are capable of bridging large regions with repetitive elements,

[9] According to the information made available by Roche reads can now reach sizes up to 1kbp (http://454.com/products/gs-flx-system/index.asp).

providing a valuable method for informative mapping of the genome.

The other type, paired-end reads, are generated by fragmenting the target sequence into segments typically of 500 base pairs or less, and sequencing both ends of the segment. These possess the same difficulties associated with the mate-pair reads generation, but provide higher resolution than these, and are essential for resolving the complexities created by some of the shorter repeats, or reads in the adjacency of repeat regions.

For both mate paired and paired-end reads, different insert-size libraries can be created to cover different levels of resolution, improving the mapping of the reads at various scales, and helping reduce the number of gaps between contigs.

The assembly issues however are not limited to genome complexity. To some extent polymorphisms, i.e. SNPs, copy number variations, and polyploidy, add some computational complexity to the already onerous assembly task. This follows from two different causes. The first is related to the way repeats are recognized in the assembly step, and the second to the confounding effects emerging from the presence of reads with errors.

Correctly addressing polymorphism detection can be troubled by the fact that compositionally similar reads can either result from the oversampling of one same region, or from different regions across the genome, which depending on the scale it is looked at, may look the same. Because heuristic approaches are often employed to perform an assembly, certain shortcuts end up being used, which can lead to some undesirable consequences. An example is the filter used by some assemblers to identify and afterwards exclude repeats. This is commonly done by the detection of reads with an average depth of coverage above the background levels across the genome (Schatz et al., 2010; Miller et al., 2010).

While often this repeat filtering can be desirable, there are cases in which it is not, such as gene duplications, which often accumulate different polymorphisms, or in the case of polyploidy, where a multitude of different alleles may be present. Indeed, failure to capture the variation between these polymorphic regions, may erroneously lead the assembler to deduce that it is in the presence of a repeat, excluding the reads involved (Fig. 1.3). This problem, present in the assembly of individual genomes, can also be observed at the taxa level when dealing with, for example, metagenomic samples. As an example, in a particular study, metagenomic samples recovered from several locations of the ocean, were sequenced and shown to have uneven levels of coverage across the reads. Reads with excess of depth of coverage were then assumed, by the assembler, to correspond to repeats, leading to their removal from the dataset. Only posteriorly with the re-identification of some of the reads as belonging to some of the most commonly occurring members of the metagenomic population, was the erroneous exclusion noticed and reversed (Venter et al., 2004). In this case, due to the filter's stringency, repeat-induced misassemblies were avoided at the cost of the exclusion of an important part of the samples in the final assembly (Pop, 2009). This highlights the need to carefully consider all parameters of a filter, by taking into account both the purpose of the experiment and the characteristics of the data. Alternatively, instead of relying on heuristic filters, one could choose to employ filters which would decide to take action, or not, by assigning probabilities to different events based on a probabilistic model, and deciding upon these.

The second issue related to distinction of polymorphisms from errors, arises from the presence of

erroneous reads, which in some instances can be difficult to distinguish from true polymorphisms. These reads may cause breakpoints in contigs whenever they diverge from the consensus, in the very same way a true polymorphism would. The essence of the problem lies in the lack of enough depth of coverage from reads from a particular polymorphic genomic region, which fail to thoroughly represent one or more of the polymorphisms present. This may lead the assembler to take essentially two different routes. To be conservative and dismiss these polymorphic reads as errors, or instead require a decrease in the stringency threshold required to identify polymorphic reads as being truly polymorphic, which also increases the chance of labeling errors as polymorphisms. This trade-off involved in the discovery of polymorphisms is relevant both at the intraindividual and interindividual level. In the former case if the individual happens to be polyploid, and is heterozygotic at a given locus so that two or more different alleles are present simultaneously, these alleles differ by as few as one nucleotidic base. Distinguishing the presence of a fragment with an error from the different allele variants may be confounded by lack of depth coverage of some allele variations. This poses a problem which is often solved by a compromise similar to the trade-off previously described. The same problem can be found if there is copy number variation, with the additional aggravate that the polymorphism is not exclusive to polyploid taxa, since it may happen within the same chromosome in haploid individuals. In the latter case, besides the issues present at intraindividual level, there is an extrapolation of these problems to the variation between taxa. This is particularly problematic when sequencing metagenomes. These are composed of genetic material from environmental samples, which means that a large number of taxa are concurrently sequenced. In this case often some taxa will occur with low-frequency, and also polymorphisms between different taxa will be present. To avoid filtering out low frequency fragments and true polymorphisms as errors, the filter's stringency needs to be relaxed. As in the repeats identification problem, at the time of the assembly it is then necessary to ponder well about how to approach the problem depending on each project final goals. In some cases it may pay off to be conservative. and lose the polymorphisms to end up with a clean assembly. Conversely, sometimes it is better to have a relaxed approach potentially leading to the inclusion of data with high error content, but which will guarantee the capture of the polymorphic variation.

All the previously enumerated issues engender a set of challenges that are both computationally and memory intensive for the assembler. Furthermore, the specificities related to the reads produced which vary by sequencing platform demand that the algorithm is fine-tuned to more aptly exploit the data richness.

Most of the algorithms developed so far to deal with the sequence assembly challenge fall mainly under two types, namely greedy and graph-based (the difference will be more fully explored in section 5). Since initially these were developed primarily to deal with the assembly of the longer and more reliable Sanger sequencing reads, usually generated in much smaller quantities, the original configurations used by the assemblers were inadequate to deal with the typical NGS throughput. Only more recently, with the need to better address the particular characteristics of NGS reads, were the algorithms adapted to address the voluminous short-read datasets produced by the NGS technologies.

Under many of these formulations the assembly problem has been shown to be NP-hard (non-deterministic polynomial-time hard) (Myers, 1995), a type of problem which is characterized by the lack of efficient computational solutions since it scales exponentially with the input size, which in

simplistic terms relates to a problem's tractability. However, so far the empirical results have shown that heuristics, which assume some simplifications to reduce the complexity inherent to the problem, perform well in practice.

Some of these simplifications encompass whole algorithm categories, while others are specific to certain assemblers. An example of a trade-off common to some assemblers is the exclusion of the quality values commonly produced for each read produced, due to the substantial increase in the use of CPU and RAM, with negligible effects in the assembly. Furthermore, the poor and limited currently available metrics, used to assess the assembler performance, may not be able to detect these quality issues. Such cases, where the quality of the assembly produced can be mistakenly interpreted as good, can be exemplified by assemblers which either produce large, but misassembled contigs, or those that produce extremely accurate contigs, but overly short. If in the former case the typical assembly metrics, such as the size of contigs or N50[10], will show optimal results, failing to reveal that these were misasambled, in the latter case the reads will show great concordancy and merge into unambiguous contigs, but due to their short size will be useless since the amount of information that can be extracted is minimal, and tasks such as gene annotation can hardly take place. These two examples represent both the extreme cases that occur when the algorithm is either relaxed, and fails to separate the "wheat" from the "chaff", where wheat represents the error-free reads and chaff the erroneous reads, or conservative, in which all the problematic repeats or ambiguous reads, true polymorphisms or not, are excluded. Deciding which solution represents the perfect balance, will often require more than the typical heuristic approach, and should instead be trusted to a method which incorporates a probabilistic-based model, conditional on the data available and a set of priors.

Sequence assembly success will then ultimately depend on the ability both to recognize that different datasets require different approaches which can fully account for the particular data characteristics, and understand the limitations that these and the genome complexity impose on the assembly task. These will allow for a better algorithm choice, which may be able to fulfill the goals set at the start of the experiment.

[10] This metric is the lower contig length threshold above which 50% of the assembly is contained.

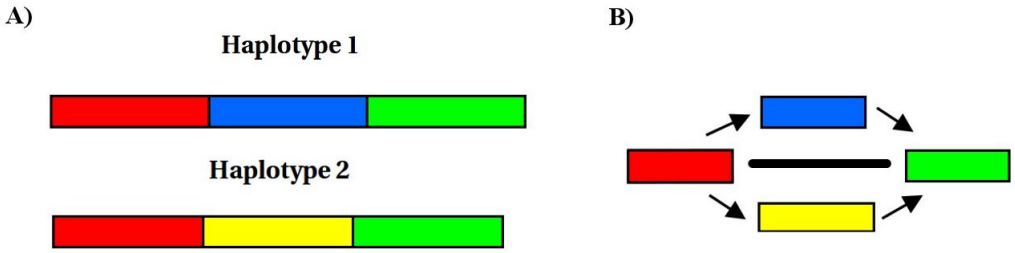


Figure 1.3: **Polymorphic sequences A)** Two haplotypes diverging in the middle section. **B)** Two true paths exist in this case, which share their ends. To identify such cases the amount of depth of coverage should be high for both versions of the polymorphism, or else one may be discarded as an error. Figure adapted from A. Policriti.

Chapter 2

A karyotypic microdissection enrichment pathway

2.1 Developing karyotypes - to culture or not to culture

To ensure the enrichment success of a specific chromosomal unit there is an implicit two fold requirement. The targeted chromosome should be clearly distinguishable, and several copies of the chromosome should be obtained.

Cytogenetically, chromosomes are not always easily identifiable. In order to distinguish them, one particular approach requires that the chromatin in the cell must first condense and arrange. This occurs during the mitotic phase of the cell cycle, when cells are replicating and preparing to divide, as a point known as metaphase. In metaphase chromosomes are deeply coiled and highly visible, and so in order to identify particular chromosomes, if any phase of the cell is to be chosen, this phase should be. Thus in order to proceed in the identification of the targeted chromosome, cells are either cultured, and the cell cycle phase is interrupted in metaphase, or alternatively, for example, cells going through metaphase can be searched for and extracted from dividing spermatogenic tissue.

Due to the burdensome features of the latter technique, cell culture is often preferred, given the greater ease to obtain sufficient chromosomal copies. However, albeit more effective for karyotype development, cell cultures even if treated with a broad range of antibiotics, are susceptible to contamination by antibiotic-resistant bacteria. Therefore this method may compromise the success of the enrichment.

To overcome the possibility of contamination two existing options are the use of the previously mentioned laborious operation with spermatogenic cells, or the use of flow cytometry[1], used for chromosomal sorting provided that the features highlighted by this approach allow the identification of the targeted chromosome.

Ultimately, the decision to culture or not to culture will have to assess if the contamination risk is

[1]It should be noted that flow cytometry may lack sensitivity to detect small chromosomes (Zhou and Hu, 2007).

sufficiently lowered by the use of a larger array of antibiotics, or if instead the other less contamination-prone methods, but more laborious and less sensititive, are to be preferred.

2.2 Microdissection

Developed approximately three decades ago by Scalenghe et al. (1981), microdissection bridges the cytogenetic and molecular genetic fields, by allowing the isolation of single cells, chromosomes, or specific recognizable genomic regions for further post-processing with molecular tools and methods.

The products of microdissection are potentially useful to many applications, namely high-throughput genomic, transcriptomic or proteomic applications, expression profiles research, generation of probes for chromosome painting, or genetic linkage map and physical map assembly (Espina et al., 2006, 2007; Rodriguez et al., 2008).

This cytogenetical technique was originally used to obtain DNA from single bands on *Drosophila melanogaster* polytene chromosomes (Scalenghe et al., 1981) and later from easily identifiable human chromosomes (Bates et al., 1986). However, mainly due to the fact that it was a laborously manual task, and good identifying techniques were lacking, it was only in subsequent years that it began to be more extensively used. This is linked to the development of an handful of techniques such as flow sorting and FISH, which allow the identification of the less easily distinguished chromosomes (Lüdecke et al., 1989; Senger et al., 1990; Meltzer et al., 1992; Yu et al., 1992), and subsequent development of technological improvements, such as laser dissection (Métézeau et al., 1993).

Currently microdissection can be divided into two major categories: manual microdissection, and the more recent and precise method, laser manipulated microdissection (LMD). The latter can be further divided into Laser Capture Microdissection (LCM) (Emmert-Buck et al., 1996; Simone et al., 1998; Lawrie et al., 2001), and Laser Microbeam Microdissection (LMM) (Böhm et al., 1997; Schütze and Lahr, 1998), where the second shows several advantages, including speed and efficiency.

The advantages of LMM over LCM can be further summarized in a couple of points. These are the choice of a ultraviolet (UV) laser over an infra-red (IR) laser, allowing for a more precise focus, the non-contact nature of the system, i.e. the target is ejected without touching any contaminated surface, the possibility of ablating unwanted tissue, and the fact that it avoids the possible molecular modifications which occur by either heating or cooling the thermoplastic membrane that contains the sample. Regarding the very last point, it is possible because the very precise pulse of UV laser can “draw” around the target cell, or cell structures. The cut target then falls straightly onto the eppendorf by the effect of gravity (Curran and Murray, 2005).

As in every other method, it is not without its own pitfalls. The disadvantages of LMD method are essentially connected to the high-cost associated to the equipment and consumables, along with the fact that it is laborious and time-consuming to retrieve the target region. Since often protocols

require several microdissections to be performed in order to gather enough genomic material, the time and exposure involved make it a process also prone to contamination (Curran and Murray, 2005). In this sense the amount of material microdissected should be kept to a minimum, and the step as time efficient as possible, so that chance contamination events can be minimized (Zhou and Hu, 2007).

Given that this is not always possible, the exclusion of the surrouding cytoplasm, which can be contaminated, may be key. In that sense, a further measure to avoid contamination was proposed by (Hu et. al 2003), in a paper describing a modification to the manual microdissection application. This proposal suggests that the target chromosome should be put in a drop of 50% ethanol. Because the surface tension present in a 50% ethanol drop is weaker than the water, the microdissected tissue will then adhere to the tip of the glass needle without any cytoplasm, and so it will not enter the drop when the tip of the glass needle is removed.

The success of the steps that follow the microdissection will largely depend on how successful this method is, both on retrieving the chromosomes, and avoiding undesirable contamination.

If these goals are achieved at a consistent level, microdissection, by effectively permitting the isolation of a particular chromosome for further processing, with little to no previous knowledge abouts its composition, can be seen an invaluable tool for any pathway which relies upon the enrichment of a particular genomic region as large as one chromosome.

2.3 Whole genome amplification (WGA) techniques

Whole genome amplification is a non-specific amplification technique which endeavours to amplify the totality of a genome or part thereof, such as whole or partial chromosomes. It does so by usage of random primers with or without partially pre-designed templates, and adaptor linkage to fragmented sequences, produced by shearing the genome, a step whose goal and issues were briefly touched in 1.

Presently three techniques are described in the literature. Linker adapter PCR (Lüdecke et al., 1989), the oldest method of the three, involves the digestion of the DNA target sequence with restriction enzymes, and linkage of adapters to its ends. These adapters have a specific sequence, complementary to that of the primers, and so each fragment previously created with adapters linked will be theoretically amplified.

A slightly more complex technique is primer extension pre-amplification (PEP-PCR). This technique involves the design of a set of random hexamers. These are used to prime the DNA template at several regions of the target sequence (Zhang et al., 1992). Subsequently, in order for the amplification of the template DNA to happen, it is subjected to thermal cycles with very low annealing temperatures, and an extra set of at least 50 cycles. Although not bias free, since the random hexamers used show non-uniform annealing and extension, the bias can be partially alleviated by using multiple displacement amplification (MDA) (Dean et al., 2001). This type of amplification can substantially improve PEP-PCR by using a mesophilic, highly processive DNA polymerase, named *phi29*. This attenuates the problems which otherwise arise when the complimentary versions of the hexamer sequences in the template sequence are either rare, or too sparsely distributed. These improvements can be seen not only at the fragment length level, with amplified segments presenting sizes ranging from 10kb to

50kb, but also in the representation indices of the amplified fragments.

The third technique is degenerative oligonucleotide primer PCR (DOP-PCR) (Telenius et al., 1992). This technique, as the name suggests, uses primers with degenerated regions. These regions can be found inbetween the 3'-end random sequence, which is intended to anneal evenly throughout the DNA sequence target, and the partially fixed 5'-end sequence. After some PCR cycles where the primers anneal to the DNA template through the 3'-end random sequence and extension occurs, the temperature is raised to 30 degree celsius and a second amplification step takes place. During this second amplification, set at higher temperatures so that the annealing occurs with more specificity, a new set of primers are designed to anneal onto the 5'-end fixed sequence now present at the end of some amplified fragments. If no problems occur, not only should the sequence be fairly covered, but additionally the oligonucleotides should have abstained from annealing with each other.

Chapter 3

Choosing a target for the trial

3.1 Microchromosomes in birds and reptiles - ease of microdissection

Microchromosomes, as the name implies, are chromosomes of reduced size. They have been found in a diverse array of vertebrate taxa, such as reptiles, birds, fishes and amphibians, but are seemingly missing from the very compartmentalized mammalian genome (Fillon, 1998), with a few known exceptions, such as bats (Oh, 1975). Their presence is particularly noted in birds and non-avian reptiles, being ubiquitously present in the former group, and prevalent in most species of the latter group (Burt, 2002).

Despite their small size (or perhaps because of it), these genomic units have been shown to have an overall gene density larger than that present in macrochromosomes, encoding for 50% of the genes, while accounting for only 25% of the genome (Burt, 2002), being particularly gene rich in chickens (Smith et al., 2000), and possessing high recombination rates which surpass those present in macrochromosomes (Chelysheva et al., 1990; Rodionov et al., 1992).

Moreover, gene mapping and sequence comparison performed between chicken microchromosomes and other vertebrate genomes revealed the presence of conserved synteny between these taxa. This fact partially supports the hypothesis that more than half, if not all, of the chicken microchromosomes may represent ancestral syntenies, and additionally that these are the product of chromosome fission (Burt, 2002).

A more recent comparison between the chicken genome and the recently sequenced and assembled *Anolis carolinensis* lizard's draft genome, revealed at least 259 syntenic blocks, as in consecutive syntenic anchors, with the same order, orientation and spacing, between the chicken and the lizard, and that the microchromosomes of the two taxa are exclusively syntenic (Alföldi et al., 2011). This is not surpring given the fact that a similar comparison between the chicken and a *Xenopus* anuran amphibean (Hellsten et al., 2010) showed that only a small number of rearrangements have taken place in the past 280 million years that separate these two taxa.

The sequencing of three avian genomes and one lizard genome have contributted to a better comprehension of the intrincacies of the reptilian, and more generally the amniote karyotypic evolution, whose study is currently constrained by lack of enough data. However, this information still falls short if the ultimate goal is to have a clear and deep understanding of the evolution history of these taxa. By focusing on microchrosomes, genomic units shown to be rich in genes and overall filled with informative content, it is possible to yield invaluable data from entire linkage-groups. More in-depth evolutionary studies can then take place, not only in a more frugal fashion but also with better perspectives of successfully extracting useful information from the data, specially if compared to more ambitious endeavours such as those contemplating whole genomes. More to the point, due to the smaller size that these chromosomes display (on average 12 Mbp in the chicken genome (Axelsson et al., 2005), ten times less the size of its macrochromosomes (Rodionov, 1996)), a smaller investment is required to ensure better breadth and depth-wise chromosome coverage. In addition the current availability of platforms which can produce up to one million reads of throughput with a mean size of 400 bp, which can represent 30 fold coverage in the case of a macrochromosome, should considerably simplify the assembly step.

Lastly, considering the fact that the lizards microchromosomes are still understudied when compared to those of birds, any input would prove important to infer both their origin and evolution, by allowing comparative studies with other taxa to be undertaken.

3.2 Lacertids: a well-studied group with no sex chromosome markers and micro sex chromosomes

Lacertids, with over 300 species presently described, ascribe to a total of 37 genera, and aggregate a diverse group of lizards commonly known as the true lizards. Distributed across Eurasia and Africa, they show a multiplicity of reproductive and sex-related mechanisms and modes, akin to that commonly present throughout the more encompassing "parent" squamate order, and show overall striking levels of diversity in terms of sex and reproduction modes (Arnold et al., 2007; Pavlicev and Mayer, 2009).

Despite the overwhelming level of diversity present in Squamata reptiles which is expressed by the presence of genetic (GSD) and/or environmental sex determination (ESD), possession of a multitude of sex chromosome systems, both female and male heterogamety with variations on sex chromosome number (ZZ/ZW, XX/XY, ZZ/ZZW), several modes of reproduction (viviparity, oviparity, ovoviviparity), and even parthenogenesis (Arnold et al., 2007), in the lacertidae family so far no ESD cases, nor male heterogamety, have been confidently observed. Thus, this group can be described as being more conserved than taxa from other families belonging to Squamata.

On lizards, which are contained within Squamata reptiles and contain the lacertidae, all of the above mentioned mechanisms of sex determination are said to have evolved multiple times. Indeed, empirical data shows that generally there is a lack of clear phylogenetic clustering among taxa which show to be male homogametic or heterogametic GSD, and/or among those shown to express ESD [1] (Ezaz et al., 2009). Additionally, so far the study of sex chromosomes in lizards has shown that

[1] However, an association between ESD and female heterogamety has been said to exist within families, for which

these possess different morphologies and degrees of degradation, further supporting the hypothesis of multiple and independent sex chromosomes origins. This fact may complicate, or even make it impossible, to compare the same sex chromosome among different lizard taxa, since it is truly possible that these might not share any homology.

Considering their diversity and rich evolutionary history, it is not surprising then that lizards are one of the most widely studied groups in nature. Nonetheless, those studying lacertids are still confronted with a lack of sex chromosome markers. These molecular markers are most prized for evolutionary biology studies, mainly due to the huge impact and extensive role sex chromosomes typically have in evolutionary processes such as speciation.

Karyotypically the greater part of lacertids have 36 macrochromosomes and two microchromosomes (Matthey, 1931, 1939b; Darevsky, 1966; Dallai and Baroni Urbani, 1967; Arronet, 1968; Kupryanova, 1969; Kupryanova and Arronet, 1969; Orlova and Orlov, 1969; Chevalier, 1969)), with exceptions such as *T. lepida* (Matthey, 1939a), and *L. strtgata* (Orlova and Orlov, 1969), both with 38. The readily available modern cytogenetic techniques, which allow the identification and isolation of most individual chromosomes or chromosomal regions, and the small size of the sexual chromosomes in this lizard family, create the conditions to undertake the enrichment of these crucial chromosomal units, warranting their study.

3.3 The interest in sex chromosome in evolutionary studies

Sex chromosomes are particularly conspicuous chromosomal units which show an ability for sex-determination and an important assymetrical distribution on different sexes. The asymmetrical distribution of sex chromosomes has profound implications on the evolutionary rates, due, not only to the fact that chromosomes end up spending disproportional amount of evolutionary time in one sex, but also because it may result in hemizygosity.

In accordance to what is stated by the former argument, sex chromosomes show smaller population effective sizes when compared to those observed in autosomes (Schaffner, 2004; Vicoso and Charlesworth, 2006). One implication of this smaller effective population size is that sex chromosomes are in a sense more exposed to genetic drift, so that the stochastic events that govern their evolution will appear to have stronger and faster effects.

Hemizygosity, however, shows an equally important role in shaping evolution. In the individual where the sex chromosomes are found to be hemizygous, such as the male in humans or a female *Eremias velox* lizard, its genes even if possessing the recessive version, will often be expressed. This is contrary to what happens in autosomes where two recessive copies or more (in the case of polyploidy) may be required. That is, for all intents and purposes the totality or most part of the loci present in these chromosomes are expressed, independently of being recessive or not. In this peculiar scenario natural selection can more readily target recessive advantageous mutations, causing their frequencies to more rapidly rise, while more easily screening deleterious mutations (Vicoso and Charlesworth, 2006). This "unprotected" exposure to natural selection results in a faster and more effective evolution of sex chromosomes relatively to autosomes.

there are also exceptions such as the Gekkonidae family.

Furthermore, in the assembly context, the presence of hemizigosity in chromosomes can be regarded as useful. This is due mainly to the fact that only a single allele for all loci will generally be present. In this sense confounding effects that often arise, from the presence of multiple polymorphic genomic regions on homologous chromosomes and the difficulty of distinguishing these from errors, do not represent a problem in the assembly of hemizygous chromosomes. The assembly effort in reconstructing these chromosomes can then be redirected to deal with duplicated and rearranged genomic regions. As a result the synteny and genomic composition inferred in these chromosomes often ends up being more reliable than those obtained when considering chromosomes that show up in more than one copy.

Further substantiating the hypothesis that sex chromosomes are preponderant in adaptive evolution, two evolutionary rules have been advanced as a result of empirical results. The first is Haldane's rule, initially proposed in 1922, after the observation that in the case that a hybrid seemingly healthy is found to be sterile, there is a good chance that it will be the heterogametic sex (Haldane, 1922). The second, the large X-effect (Coyne and Orr, 1989), also termed Coyne's Rule (Turelli and Moyle, 2007), is the observation that, where hybrid sterility is present, the X chromosomes effects seems to play a larger role than other chromosomes. The first rule is now also said to account equally for the disproportional rate of hybrid inviability in the heterogametic sex. Regarding the latter, it has been shown that the Z sex chromosome, from the ZZ/ZW system, also has such a disproportional effect in speciation[2]. This is perhaps a reason to favour the name "Coynes rule" over "large X-effect".

Haldane's empirical observations have been mainly explained by the dominance and fast-male theories. The dominance theory suggests that the alleles implicated in the reduction of hybrid fitness, leading to the emergence of hybrid incompatibilities, will on average be partially recessive. As a consequence of being recessive these alleles will then be fully expressed in heterogametic hybrids, if they happen to be (Z or X)-linked, whereas in homogametic hybrids their expression will be either non-existent or negligible. The fast-male theory states that in the XX/XY sex system, incompatibility factors said to cause male hybrid sterility will accumulate faster than other types of incompatibilities. The causes may be due to sexual selection acting on male-specific genes, or low tolerance to perturbation during spermatogenesis.

The large X-effect is related to the disproportional contribution of the homogametic sex chromosomes to the hybrid heterogametic incompatibilities. It was first shown to happen in *Drosophila*, in which it was noticeable that recessive alleles accounting for hybrid male sterility were largely concentrated in the X chromosome. By backcrossing species hybrids, with introduction of one species's X into another species, and vice versa, a noticeable effect was perceived on hybrid fitness as compared to that observed in autosomes. Since the previous observation, the same effect has also been revealed to be present in numerous other species (Coyne and Orr, 1989; Coyne, 1992), and interestingly it was demonstrated, both theoretically and empirically (Ellegren, 2009), that it may have a stronger effect on the Z chromosome under the good genes (Zahavi, 1975) and Fisher's runaway (Fisher, 1930) models of sexual selection. The good genes model hypothesizes that if females show a preference for males with a particular trait, which might not be related to male fitness, that trait will be advantageous by association with the underlying females preference. This leads to an increasingly stronger sexual selection of such traits in a population towards extreme values. The growing trend may continue until the balance between having the advantageous trait, and the disadvantages it might entail, is

[2]This effect has been conveniently dubbed the large Z-effect (Ellegren, 2009).

disrupted, or the relevant genetic variation becomes exhausted. In essence this model relies upon, and will tend to increase, association between the trait, and the preference for this trait, for example by reducing recombination between them (which breaks down associations). Fischer runaway model suggests that female preference will go towards males showing traits that may be indicative of some advantage providing better fitness. As an example, brighter colours in males may be an indicator of better health. The consequences of this selective preference are the same as the ones described for the former model.

In the context of the large Z-effect this could be expressed by considering a Z-linked gene which determines the female preference. In this case, and since females are the homogametic sex in the ZZ/ZW system, the females possessing this preference will pass it to at least half of the male progeny, which will additionally have the gene deemed as preferable by females (Ellegren, 2009). It can be suggested then that bright and visible displays will be more prone to appear in ZZ/ZW systems (Albert and Otto, 2005). Interestingly this former hypothesis is supported by empirical data in diverse taxa (Hastings, 1994; Prowell, 1998; Reinhold, 1998; Volff and Schartl, 2001; Iyengar et al., 2002; Mank et al., 2006).

The precise role that these chromosomes play in evolution, and how they exactly came to be, remain questions requiring further research. Answers to these questions will results in a better understanding of the evolutionary processes, and of the diversity that results from them.

Project Goal

This project is a proof-of-concept trial whose primary goal was to develop a fast and economical pipeline to obtain genomic information for individual chromosomes. This involves combining cytogenetic and molecular tools and methods to select and enrich for a particular chromosome, followed by its sequencing and a final assembly step performed *in silico*. In particular, to show the concept feasiblity the goal would then be to microdissect 12 W sex microchromosomes from a blood cell culture of a female lacertid, amplify the amount of available genomic material by employing a whole-genome amplification protocol, and sequence with NGS to then proceed to its assembly. The sequencing output with this kind of sequencing technology should provide enough depth and breadth of coverage to enable the computational assembly of the chromosome in its entirety, through a less laborious and costly method compared to older sequencing methods. Finally, to validate the assembly produced, some of the assembled contigs should be put through lab validation to see if they were truly well assembled and if they represent the lacertid.

Methods

Chapter 4

The trial pipeline - from *Eremias velox* to 9x10^5 NGS reads

4.1 *Lacerta schreiberi*: failed leucocyte culture

With the purpose of selecting a lacertid with a ZZ/ZW sex system, for W sex chromosome enrichment, blood samples were taken from a female *Lacerta schreiberi*. In this species, as in all studied lacertids, the female is the heterogametic sex, thus bearing both which is the heterogametic sex thus bearing both the Z and W sex chromosomes.

Leucocyte cell culture was performed, but good metaphases were not obtained. This step was performed by Martina Pokorná, from Charles University in Prague.

4.2 *Eremias velox*: successful leucocyte culture

In virtue of the previously failed leucocyte culture with a *Lacerta schreiberi* lizard, another lacertid was selected. The species chosen was *Eremias velox*, since it is known to possess a ZZ/ZW system, and additionally, previous lab work showed that good metaphase suspensions could be obtained for both sexes in this lizard.

A female *Eremias velox* was then selected, whence blood samples were taken.

For leukocyte cultivation, peripheral blood was cultured at 30°C for a week in T 199 medium (Sigma-Aldrich), enriched with 10% fetal bovine serum (Baria), 0.5% antibioticantimycotic solution (Sigma-Aldrich), 1% canamycin (Sigma-Aldrich), 0.2% phytohaemaglutinin (Biomedica), and 1% lipopolysacharide (Sigma-Aldrich).

Subsequently, for prospective visual identification of the lizard's chromosomes, metaphase chromosome spreads were prepared from cultures of whole blood following the protocols described by Ezaz et al. (2005) with slight modifications.

Martina Pokorná carried out all the previous steps.

4.3 C banding

Eremias velox sex chromosomes are homomorphic microchrosomes, hence indistinguishable at the microscope.

In order to ascertain which sex chromosome is which, several staining techniques were tested. Ultimately C banding was found to be capable of distinguishing the W and Z chromosomes present in the female *Eremias velox* lizard, in this case by staining the W chromosome, while leaving the Z unstained.

C-banding was carried out following the method described by Pokorná et al. (2010). Both the previous experimentations and the C banding were performed by Martina Pokorná.

4.4 Microdissection of 16 exemplars of the W chromosome

To isolate the microchrosome identified by the C banding technique, the metaphase suspension of cultured leucocyte cells was first dropped onto a special sterile membrane. Laser microdissection was then performed by Martina Pokorná using the Olympus laser microdisector. The isolated microchromosomes were ejected into a drop of TE buffer located on the eppendorf tube, then centrifuged, and lastly placed in a freezer.

The small size of the chromosomes and the explicit desire to enrich them, warrants the requirement of microdissecting a total of 16 W sex microchromosomes. This is the same number of chromosomes required to make FISH probes from microdissected chromosomes, and is in accordance to the standard requirement of 10 to 20 copies of microdissected chromosomes to perform amplification (Zimmer et al., 1997; Ráb et al., 2008; Henning et al., 2008).

4.5 WGA

In order to decide which method to use it is important to take into consideration the possible drawbacks that may be associated with the techniques chosen. In the context of this trial a kit which performs DOP-PCR was selected. The kit in question uses the linkage of 30 bp adaptors to the sequence in order to prepare it for amplification. This can be, depending on the aim of the experiment, possibly problematic if the intended next steps are NGS and assembly. The reason why this choice could be problematic lies on the necessity of adaptor inclusion on the DNA fragments. Given that the NGS technologies produce reads of shorter size compared to those generated by Sanger sequencing, the inclusion of these 30 bp adaptors makes up for a huge portion of the sequence read generated, insofar the amount of information retrieved is substantially diminished. In this sense, Roche's 454 sequencing technology known to produce at the time of the experiment the longest reads from all the NGS platforms, was selected to ameliorate the amount of sequencing read portion lost

to adaptors. While it generates less reads than other rival platforms, producing 10^6 reads in a single run, the read length can go up to 700 bp, averaging at 400 bp. This can yield a total of 500 million base pairs. Such features are ideal for sex microchromosome sequencing, which can be riddled with complex regions that should be contained within single reads if possible. This combination of longer reads and the possibility to easily attain 20x coverage should allow us to more effectively tackle such this trial's assembly challenge and justify the platform choice.

Whole genome amplification was performed with Paula Campos in a clean lab at Niels Bohr Institute - University of Copenhagen, Denmark.

The kit used was the GenomePlex Single Cell Whole Genome Amplification Kit (WGA4; Sigma Aldrich), which claims to perform a million-fold amplification. Especially designed for single cells, it improves the chances of generating a representative genomic amplification, from a minute quantity of starting DNA sample as is the case of the microdissected sample.

Amplification of the microchromosomes was performed following the manufacturer's instructions (Sigma Aldrich). The protocol includes a cell lysis step, which includes not only the lysis step but also the fragmentation of the sample's genome, followed by an isothermal library preparation, using a primer which provides good coverage at low template, while refraining from self-annealing, and finally the actual amplification.

The whole genome amplification yield was then tested in a electrophoresis run on a 1.5% agarose gel.

4.6 NGS

Roche's 454 sequencing platform was selected for the sequencing step, using Titanium FLX series reagents. The amplified fragments produced by the WGA4 kit were prepared and processed according to the 454 FLX Titanium Library construction kit and protocol (Part et al., 2010).

Chapter 5

From 9x10^5 NGS reads to contigs: Alternative assembly approaches

The output of the sequencing step consisted of files containing the sequencing reads and their respective quality per base scores. The sequencing reads and quality values files possess for each entry an identifier, and furthermore, the sequencing reads have both upper-case and lower-case characters, where the lower-case characters correspond to nucleotide bases which were deemed of low quality by the sequencing platform software.

To reconstruct the original sequence, and then extract the most information out of it, reads have first to be pre-processed, and only then should they be assembled.

The pre-processing step is simply the procedure of trimming flanking base pairs. It is commonly applied to regions of the reads which aren't part of the original sequence. In this particular experiment, reads were composed by a key[1], a four character based string that precedes every read; the amplification and sequencing adaptors; the multiplex identifiers; and the sample DNA sequence. The latter are commonly used to distinguish concurrently sequenced samples, that may need to be distinguished in the same sequencing run. Additionally, regions which have nucleotides with low quality values may also be eligible for trimming. While parts of the reads that are extraneous relatively to the original sequence should ideally be trimmed, the decision to exclude nucleotides with low quality values is optional. Sufficient depth of coverage may provide enough power to decide a posteriori exactly what base pairs are untrustworthy, effectively allowing us to incorporate the most base pairs possible with minimum uncertainty about their reliability.

An equally important step which precedes the assembly, is the choice of the assembler. As described earlier in Section 1, the current NGS data, with different characteristic error profiles, biases, and particular features, require that the algorithm chosen to process them ought to be fine-tuned for all these intrinsic read-related details, as well as optimized for speed and memory usage, taking into account the final goal of the project.

With this in mind, on this trial two distinct assemblers were used. Their choice follows directly from a

[1] This key is the same for every read sequenced on the platform.

reasoned deliberation based on the previously mentioned variables and their respective weight. First, Newbler, Roche 454 Life Sciences official assembler (Margulies et. al, 2005), was used. Subsequently an algorithm developed in the Mathematica programming environment (unpublished) by Stuart J.E. Baird was used with the intent of overcoming the limitations which Newbler came across, and as a way to better control and understand the flow of the assembly. These algorithms will be briefly described and their results will be compared in Section 6.

Newbler version 2.3 was primarily chosen based on its algorithm suitability to deal both with long reads, and 454 sequencing reads specific error-profiles, but also due to the valuable inclusion of supplementary read related information. This information, stored in the read flowgrams, allows for homopolymer error correction to be performed more effectively in the assembly step. It should be noted that this type of error is recognized as one of the most prevalent issues associated with the 454 sequencing platform generated reads.

The algorithm that best characterizes Newbler is overlap-layout-consensus (OLC). It is based on a string graph approach, originally developed for Sanger sequencing reads, which is optimized for large genomes, but also deemed to be expedient for very short reads or long reads of small genomes (Chaisson and Pevzner, 2008; Miller et al., 2010). It formulates the assembly problem as a graph, where nodes correspond to sequencing reads, edges to the overlap between the reads, and where each read must be traversed exactly once. Such a path through the graph is a Hamiltonian path, i.e. every node of the graph is used exactly once. This last instance is known to lead to a NP-Hard optimization problem, and thus heuristic strategies are frequently used to solve it in polynomial time.

These heuristic solutions include the removal of transitive edges[2], collapse of overlaps with no conflicting edges, possibility of mate-pair reads employment for coupling, and ordering of the contigs.

More generally, the overlap-layout-consensus algorithm can be separated into three stages: (1) In the overlap stage, overlaps are computed in an all-against-all pairwise read comparison, and a graph structure is created. In this step, seed & extend heuristic algorithms may be used to ameliorate the process. These work by dividing the reads into k-mers (subsets of k length present in the read), which are then used to find candidates with a similar amount of the same k-mers; (2) In the layout stage, the graph is simplified by removing redundant information (collapse of contained contigs), and their proximate order in the genome is inferred; (3) In the consensus stage, multiple sequence alignment (MSA) takes place, where all reads are aligned to the contigs, and the consensus sequences and their layout is determined (Miller et al., 2010).

Peculiarly, Newbler implements this OLC strategy twice. In the first cycle, reads are divided in two classes, long and short, and unequivocal overlaps (the default threshold is 40 base pairs) between reads are searched, so that unitigs (uniquely assemblable contigs) can be generated. For efficiency purposes Newbler makes use of the heuristical seed & extend algorithm previously mentioned, and by default produces 16-mers seeds from the reads, each starting 12 base pairs upstream of the previous 16-mer. These trustworthy unitigs are then used as seeds, and compared in a pair-wise fashion, in

[2]Transitive edges are those that occur between two nodes which could otherwise be connected by irreducible non-transitive edges provided there are extra nodes between them. Removal of these edges will not then result in loss of information since the same information is present in the irreducible edges.

order to create a contig graph from the overlap of the unitigs. Following this step, a disentangling phase occurs to simplify the graph. In this phase, if it is the case that unitigs have their prefix and suffix aligned to different contigs, these may be split, with their reads being also split across different contigs, which can be a consequence of having acquired chimeric reads or representing part of a repeat region. Additionally, in the MSA step, Newbler aligns the sequences to the obtained contigs in order to obtain a consensus. It does so not only by using the depth-coverage provided by the alignment of sequences, and their quality values, but also by computing the unitig and contig consensus in "flow space", that is, using the information stored in a flowgram file containing the signal strength correspondig to each sequenced nucleotide base.

To correctly make use of its invaluable "flow space" information, Newbler's input must consist of the exact original reads. It follows then that if any noise characters are to be trimmed by the pre-processing step (corresponding to adaptors, and optionally to low quality values), their trimming points for each sequencing read should be set in a separate file. Otherwise Newbler will initially assume that only the lower-case characters should be trimmed.

The final output of the assembly consists of files with the contig consensus and their respective quality scores, together with a file containing the multiple alignments processed during the run, and also an assembly metrics file.

With all the parameters set to default, an assembly run was performed with Newbler. The input files consisted of a binary file which contains the sequencing reads, their respective quality values per base, and the flowgram information per base, and a file containing the new trimming points. The new trimming points were obtained by running a Perl script through the file containing the reads, and determining where should the new sequencing read positions start and end, so that the non-informative portions of the reads were excluded. Additionally windows of five base base pairs flanking the sequence which showed an average quality under 20, where 20 corresponds to 99% accuracy in base call were also excluded. Lastly, and in conformity to Newbler's minimum default read length threshold of 50 base pairs for single reads, those which originally, or subsequentially to its trimming, became shorter than this threshold, were excluded.

The output of the first assembly run was compared against the National Center for Biotechnology Information (NCBI) nucleotidic database with the BLAST algorithm (Altschul, 2005). The BLAST alignment results revealed the presence of bacterial contamination in some of the contigs assembled. Therefore, reads involved in the assembly of those contigs were excluded from the assembly dataset, and the remaining reads assembled once more. The process of identification and removal of reads from contaminated contigs, and reassembly of the remainder reads was repeated until there was no evidence of bacterial contamination within the contigs. For added stringency, the BLAST alignment algorithm was used not only against the nucleotide database (blastn), but also against the protein database, which involved first translating the contigs into their putative protein sequence in all the possible six reading frames (blastx). The criteria for exclusion following the BLAST alignment was based on a maximum e-value of $1*e^{-5}$. Any alignment against the genome of the putative contaminant observed to have a value below this threshold was assumed to be the result of contamination and therefore excluded. The results of the five assembly runs performed to produce contigs free of the principal source of contamination are presented in Section 6.

As a consequence of Newbler limitations, algorithmical heuristic choices, and the presence of contamination, the output obtained from these assemblies was low in number of reads incorporated but also in terms of contigs metrics.Thus, the best course of action to achieve better results was determined to be the development of a pipeline that maximizes the amount of information used, aims for longer contigs assembly, and makes extensive use of depth of coverage to increase confidence to these. Such a pipeline was developed to address Newbler limitations, using an algorithm comprised of several sequential phases. These phases included the pre-processing step of the reads, their information-lossless compression and encoding, an iterative assembly of reads into contigs, the mapping of all the reads to the produced contigs, and finally the decoding of the reads into their original state, so that the consensus could be inferred. In the pre-processing step, regular expressions are used to identify and separate into groups each non-informative part of the read, i.e. key, adaptors and multiplex identifiers, as well as the informative portion of the sequence fragment. The list of sequences containing the informative portions of the sequences is then selected for further processing, and the others dismissed. The sequences are sorted, and re-ocurring ones counted, resulting in a new list consisting of the number of times each string is repeated, and their respective sequence.To alleviate part of the computational complexity that arises both from the presence characteristic homopolymer errors on reads produced by the 454 sequencing platform, and the length of the reads, the strings were losslessly compressed by reducing consecutive repeated characters to a single character (run length encoding, with stored run lengths).An important compression step happens at this stage. The 60 possible tuples of lengths of two, three, and four, created by the combination of each of the four characters that form the genetic alphabet {A,C,G,T}, where each character appears at most once by tuple, are encoded into 60 non-latin ASCII characters, in an one-to-one relationship. This compression not only reduces the size of the strings, but also addresses possible problems that could arise from reads containing tandem repeats of length two, three, and four, by collapsing them into a single ASCII symbol (again storing run lengths).After the completion of this compression step, an iterative process of read and contig extension follows, using a seed-based strategy to search for overlapping reads. This step first extend a picked read to the right, and involves choosing a seed with a pre-determined length, which must not have been previously incorporated in other contigs. Picking an appropriate seed length is a crucial step for the assembly to be successful. The particular seed length of 25 for this trial was established after empirically observing that it produced longer contigs more consistently. To extend contigs, a search for the longest common substring between the chosen seed and the available reads takes place. The particular way of finding reads to build contigs dictates that setting a low seed length will result in the capture of many unrelated reads since the criteria for matching is very relaxed. On the other hand, using seeds of length much larger than 25, given that the string length corresponds in its decoded and uncompressed form to a length of more than 50 characters, would result in the capture of too many strings as well, since there are more substrings to be found across larger seeds and other strings. Large seeds would then lead to several unwarranted matches, occurring solely by chance. In this regard, a second criterion, the minimum overlap between the seed and capture strings, plays also an important role, by dictating the specificity of these matches. Setting up these two parameters, comprises then a most important pursuit. Their choice should strive to find a delicate balance between the maximum level of the assembly's computational complexity allowed, and the sensibility required to capture mostly reads which are actually related. Lastly, the seed is selected from the center of a string, in order to more consistently merge the newly formed contigs into the previous built ones, since the edges are more likely to be the source of possible inconsistencies between reads. Each seed selected is then used to

look for other strings matching over more than a minimum overlap length. In this trial that threshold was set at seven characters, since that way the match will be unique enough, but not too specific to miss reads with errors. In the case that 20 strings are found to match, these are increasingly subjected to higher degrees of stringency in terms of the portion of characters they must share with the seed, so that at least 10 of the strings with the most overlap with the seed are selected, otherwise the available strings are used. In the following steps the algorithm proceeds to align the previous found strings. It produces two new strings out of the consensus of the aligned reads. One is a more conservative guess based on a threshold coverage, while the other is a more relaxed guess obtained by looking for the number of consecutive characters in each string with minimum coverage. The more conservative guess is used to check if the read is contained within an existing contig, in which case it will be merged, otherwise the alignment constitutes a new contig. This iteration goes on until every string has been checked, i.e. either already compared or part of a existing contig, and then the contigs produced in the last step are used as seeds for the left extension steps, which proceeds in the same manner. After both right and left contig extension steps are finished, the contigs generated by both extensions, which correspond to the same contig, are merged. These contigs will then be used to as the initial set of reads to repeat the read extension process, but now including only a data set composed of the reverse complement of reads not previously used to extend the contigs. The generated contigs are used to "fish" and align any string that matches at least 10 characters, a higher threshold than before since the contigs should be already reasonably extended, and there is only interest in retrieving strings that truly belong to the contig, given that these will make up the consensus. These strings are then decoded again to the original character composition, and their consensus at each previous ASCII character position is inferred.

Results and Discussion

Chapter 6

Interpreting the NGS output

The 454 high-throughput sequencing machine run produced a total of 997,462 reads with an average read length of ~288 characters. This figure was inferior to the expected 400 base pairs of average read length officially reported by 454 Life Sciences for runs performed with the 454 GS FLX sequencer using the Titanium chemistry. The length distribution of the reads can be seen in figure 6.1. This same histogram was replotted after the detection of contamination, to infer if reads mapping to the source of contamination could explain the peak around ~500 bp in the previous histogram. From the overlay of the histograms belonging to the reads mapped versus those which did not map, in addition to the original histogram, it was possible to observe that both types of reads contributed proportionally to this peak (6.2).

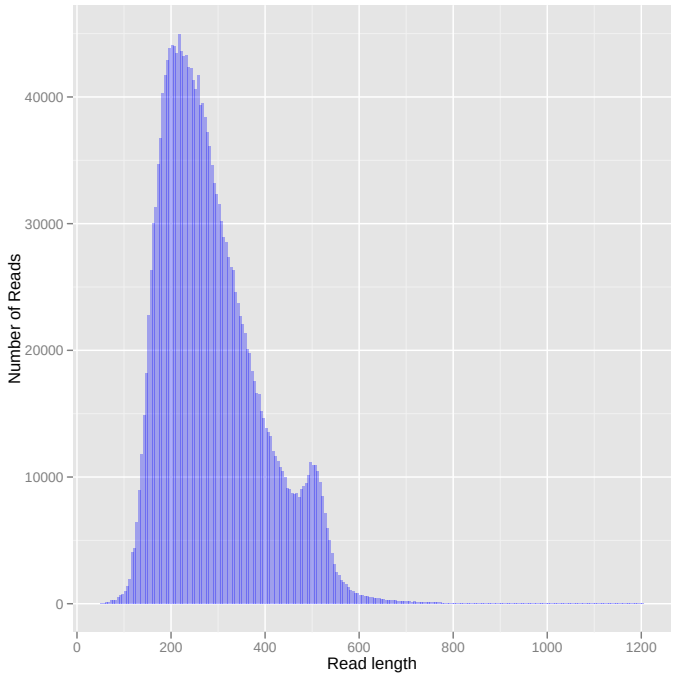


Figure 6.1: Histogram with original reads' length distribution

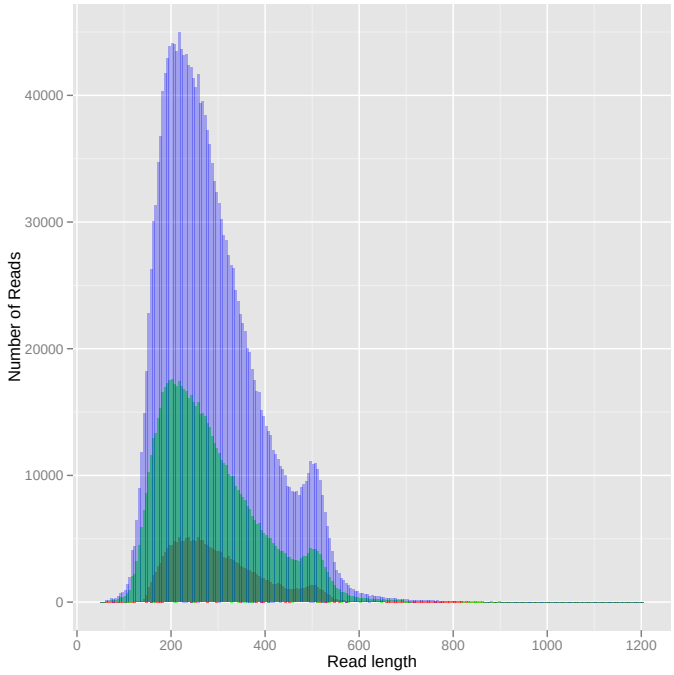


Figure 6.2: Histogram with original reads' length distribution. Overlay of the reads' length distribution histograms for reads that mapped against contamination source (dark green), those reads which did not map (lighter green), and for the initial total set of reads (blue).

The reads used in Newbler were trimmed as previously stated in section 5. A total of 269,117 (26.98%) trimmed reads were excluded due to being less than 50 characters in length, the default minimum threshold length in Newbler. The average read length after the trimming step, and excluding reads below 50 characters, was of ~180 characters, and their length distribution can be seen in figure 6.3.

Summary statistics relative to this assembly and ensuing assemblies can be found in table 6.1.

Table 6.1: Assembled Contigs Summary Statistics. Newbler was run 5 times, and the Mathematica Pipeline only once.

Assembler (# trial)	# Contigs	Mean Length	Median Length	Max Length	Min Length	N50 Length	Total(MB)
Newbler (1)	9609	377.1	289	6056	100	446	3.62
Newbler (2)	6417	340.5	261	6051	100	396	2.19
Newbler (3)	6332	336.8	260	6051	100	393	2.13
Newbler (4)	6331	336.8	260	6051	100	393	2.13
Newbler (5)	2703	305.8	252	2583	100	353	0.83
Mathematica Pipeline (1)	8405	757.0	460	10603	73	1236	6.36

The BLAST results obtained by aligning the 9609 contigs assembled with Newbler's first assembly to NCBI's nucleotide database *nt* indicated the presence of a possible bacterial contamination as seen in figure 6.4 and table 6.3. A closer inspection of the alignment results permitted us to infer that the majority of the hypothetical contamination contigs matched those of the *Stenotrophomonas maltophilia* bacterium (figure 6.5), a multi-drug resistant environmental gram negative bacterium frequently present in clinical environments.

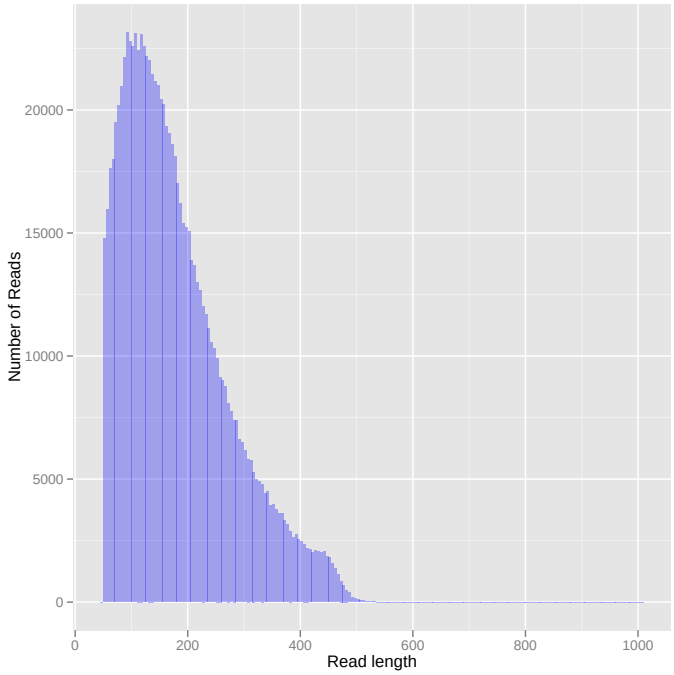


Figure 6.3: Trimmed Reads Length Distribution

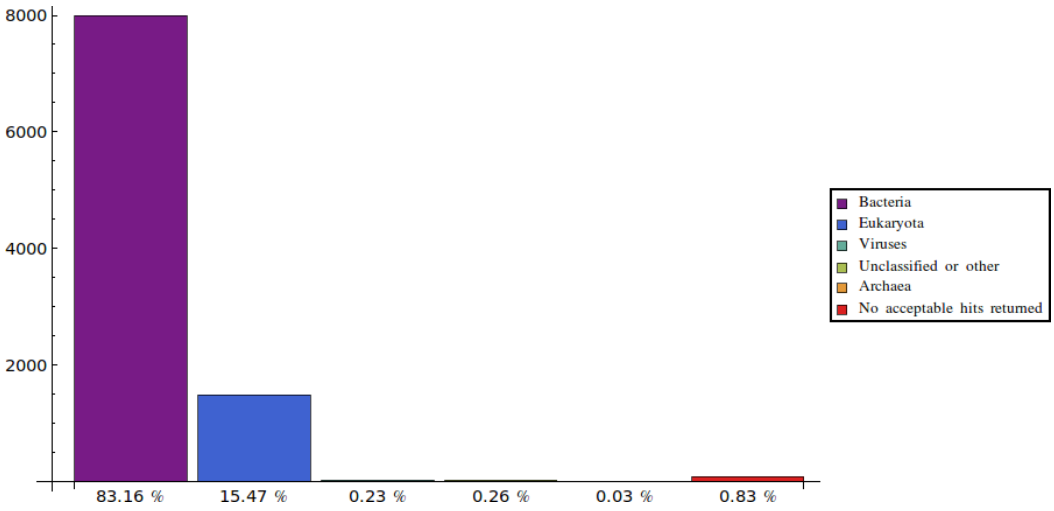


Figure 6.4: Histogram of the number of BLAST first hits for each contig by domain or match description for Newbler's assembly.

The majority proportion of bacterial contamination as indicated by the BLAST results led to the decision of performing subsequent assemblies with input datasets that exclude reads contained within contigs matching the *S. maltophilia* genome. This was done first by aligning the contigs generated in each assembly to the bacterium genome until no more contigs align to it, followed by an alignment to the non-redundant NCBI's protein database *nr* using BLAST's *blastx* algorithm, in order to remove sequences previously missed due to their higher nucleotidic divergence from the reference genome, but which could still show conservation at the protein level.

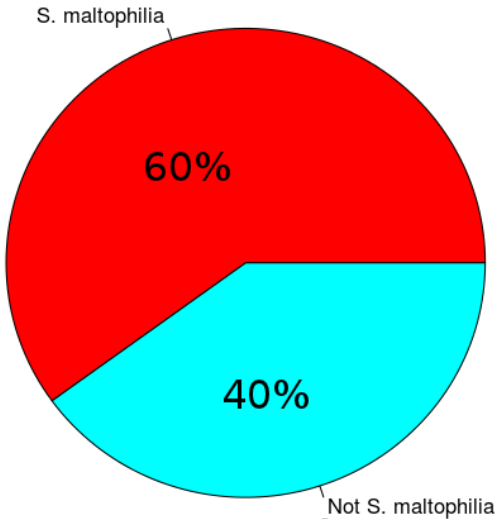


Figure 6.5: Pie chart representing the contigs from bacteria which did and did not match *S. maltophilia*. The contigs not identified as *S. maltophilia* don't belong to a particular group of bacteria.

The process of removal of reads deemed as originating from contamination and reassembly of remaining reads was performed a total of five times. By the end of the fifth and final assembly, the size of the dataset was of 364,223 reads, corresponding to 35.5% of the initial read set (see table 6.2).

Table 6.2: Number of reads used in each assembly, and their proportion relative to the original amount of reads, by assembler used, and the number of the run.

Assembler (# trial)	# input reads (% of total)
Newbler (1)	728.343 (73.0%)
Newbler (2)	501.600 (50.3%)
Newbler (3)	484.287 (48.5%)
Newbler (4)	484.241 (48.5%)
Newbler (5)	354.223 (35.5%)
Mathematica pipeline (1)	997.462 (100.0%)

In order to try to attain better assembly results, the Mathematica pipeline (MP) was used to redo the assembly in more controlled conditions. This would be achieved mainly by including reads potentially relevant to the lizard's sex microchromosome assembly, but which were excluded in the steps prior to the several Newbler' assemblies, either due to the stringent trimming applied, or based on the generated contigs' similarity to *Stenotrophomonas maltophilia*'s genome. If effective this approach should maximize the amount of sequencing reads used to perform the assembly, improve the assembled contigs length, and the degree of confidence of their assembly, by harnessing the power conferred by the increased depth of coverage.

The trimming step employed by the MP managed to include all the initial reads and further compress, without loss of information, the dataset both in size and at the read length level. The initial compression step was comprised of a 17% reduction in the total number of reads, achieved by collapsing repeated reads into unique reads, and was followed by a 25% compression attained by

run-length encoding of the read's characters, corresponding to a total of 38% compression of the initial dataset. A final 27% compression of the previous dataset was achieved by encoding every two, three and four characters possible combination into a single ASCII character, substantially alleviating the computational effort required to assemble the sequences.

Given that the latter four out of the five assemblies performed with Newbler involved exclusion of the reads identified as bacteria based on the BLAST results, which neither the first Newbler assembly nor the MP performed, the comparison between results obtained by both assemblies was only performed between the latter two, and accordingly in the following results only these are described.

Contig metrics

A comparison of the contig length distribution from both assemblies (figure 6.6) shows that the contig lengths appear to be approximately Poisson distributed, with the MP distribution having both a higher mode and a heavier tail towards longer contigs. This latter observation is reinforced by a boxplot of the contig length distribution (figure 8.1 in Supplementary Images section).

To better understand just how much of the assembly is contained within the largest contigs, the N50 metric was calculated following the definition described in the literature (Miller et al., 2010). This metric represents the lower contig length threshold above which 50% of the assembly is contained. The results for the six assemblies can be seen in table 6.1. In figure 6.7 it is shown for each respective assembly the cumulative contig length, from the largest to the smallest contig, and its N50 mark represented by the ordinal position of the contig setting the threshold. The N50 length metric, which differs from the N50 metric since it is represented by the contig length, instead of its ordinal position, can also be seen in table 6.1.

Overall the MP seems to have outperformed Newbler by attaining a N50 of 1236 bp versus Newbler's N50 of 446 bp. This means that more than half of the nucleotides assembled using the MP, are contained within contigs larger 1236 bp.

Given that the majority of the contigs map to the bacteria genome, this same metric was reanalyzed taking into account only the contigs which were shown to map to it. The results continued to show the previous trend (figure 6.8).

The apparent superiority of the MP based on the N50 metric should not however be taken at face value, given the use of a seemingly more greedy approach by the MP. This feature may make the MP more prone to erroneously adjoin unrelated reads, particularly towards the outermost contigs' coordinates. These regions are usually less covered, and thus posses a lower degree of confidence relatively to the contig's core region. The same rationale can be applied to explain MP's superior contig length metrics displayed in table 6.1. The better suitability of the MP to perform an assembly would then depend on it to display a low misassembly rate.

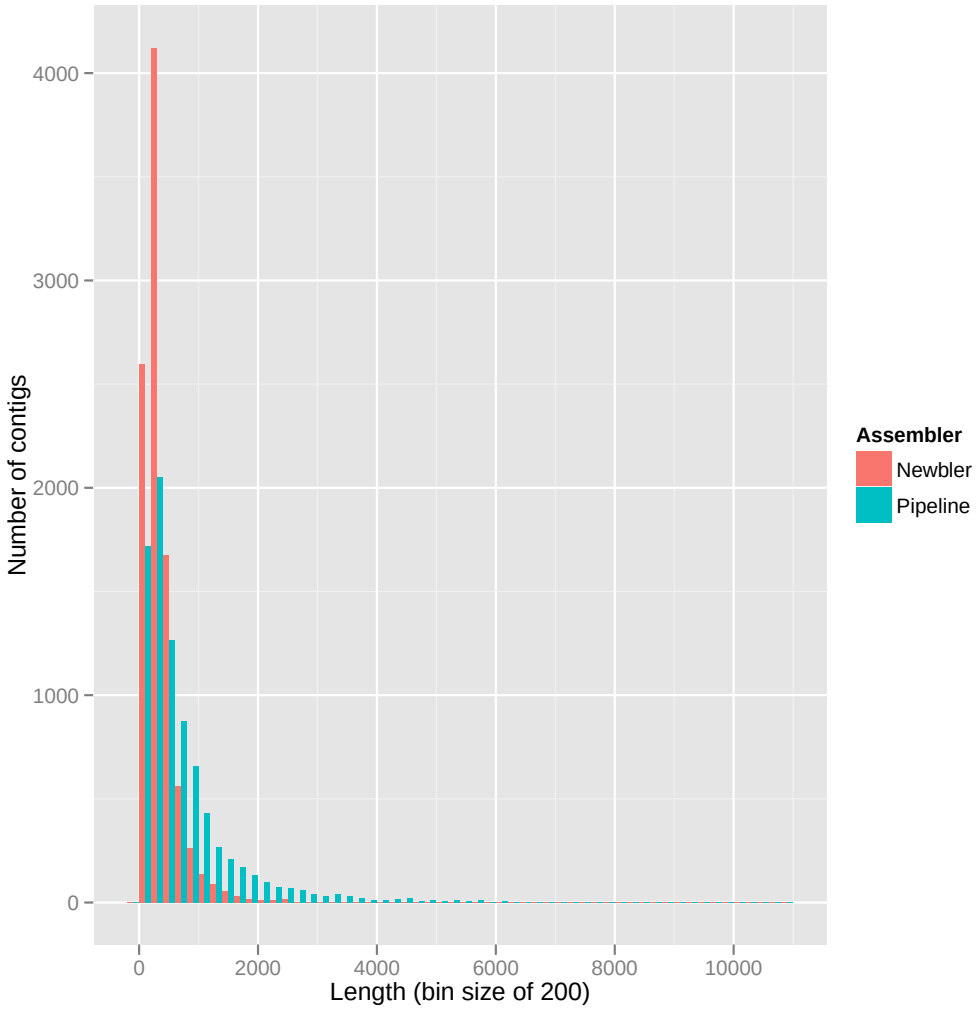


Figure 6.6: Histogram of contig length distribution for both assemblies.

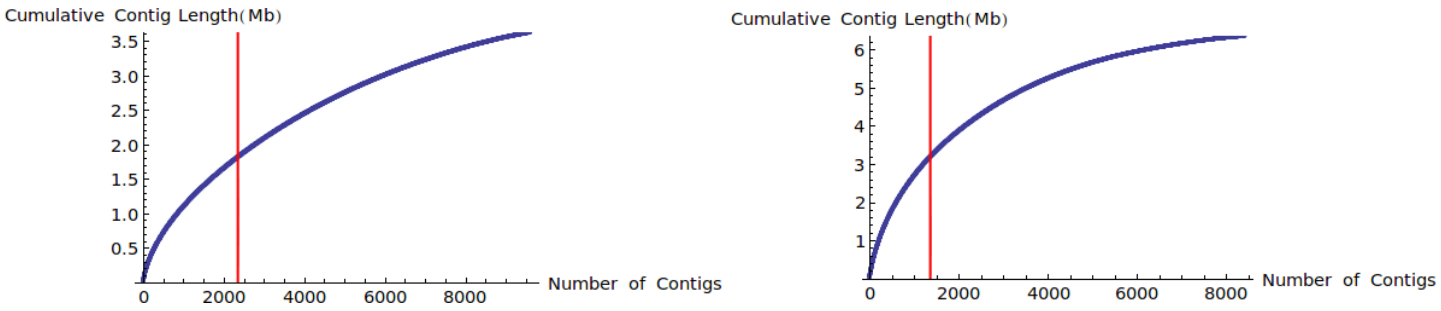


Figure 6.7: **Cumulative contig length from largest to smallest contig and N50 for Newbler(left) and MP(right) assemblies.** Red line shows to which contig the N50 metric corresponds. It is equivalent to the minimum contig length present in a set of contigs sorted by length, and the sum of their lengths correspond to at least 50% of all the nucleotide bases in an assembly.

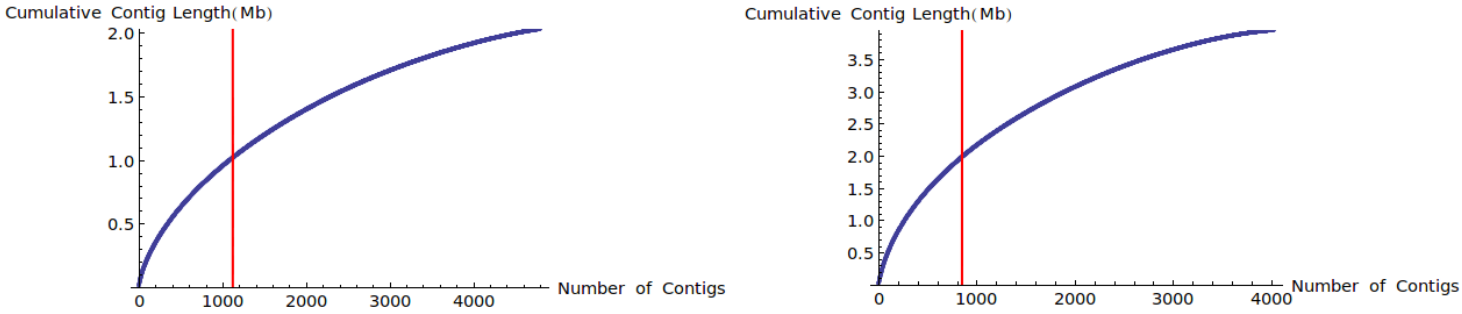


Figure 6.8: **Cumulative contig length from largest to smallest contig and N50 for Newbler(left) and MP(right) assemblies, for contigs shown to map to the** *Stenotrophomonas maltophilia***'s genome.** Red line shows to which contig the N50 metric corresponds. It is equivalent to the minimum contig length present in a set of contigs sorted by length, and the sum of their lengths correspond to at least 50% of all the nucleotide bases in an assembly.

BLAST results

By inspecting the BLAST results, it is particularly noticeable the higher portion of contigs with no acceptable hits produced by Newbler compared to those produced by the MP (table 6.3). In the MP's assembly, whose BLAST results can be seen in figure 6.9, only 12 out of the 8405 contigs (0.14%) failed to find a hit in NCBI's nucleotide database, against the 80 out of 9609 contigs (0.83%) generated by Newbler. This result can be interpreted as a measure of the MP's better efficiency to assemble reads into long and correctly assembled contigs, and/or Newbler's higher rate of contig misassembly. However, althought there should not exist a particular reason for Newbler and the MP to assemble reads from different taxa differently, another possible explanation would be that this is an indication that Newbler was able to piece together the reads corresponding to the lacertid microchromosome, since it might perform overall better. The finding of these contigs would agree well with the lack of reliable lizard's W chromosome references in NCBI's database, and the fact that it failed to align with any of the bacterial sequences. In this sense a noticeable amount of contigs with no matches would not come as a totally unexpected result. More so, considering that reads should originate from a (likely fast evolving) sex chromosome, and that reptile taxa are known for possessing pervasive high levels of diversity at that level, the chances of finding similarities with other taxa can be low.

The contig length distribution discriminated by taxa for both assemblies (histogram 6.10, and a boxplot 8.2 in the Supplementary Images section), show again that the bulk of contigs' BLAST results coincide with bacteria. Additionally, these contigs happen to be the longest. This observation goes well with the idea that since most of the reads originated from bacteria it would be more probable to see longer contigs identified as belonging to this taxa. In addition, it can be seen that while in Newbler's assembly most of all of the contigs identified as bacteria are less than 1kbp, the MP has a greater proportion, compared to Newbler, of contigs with larger sizes, albeit more than half are less than 1kbp length. Relatively to Eukaryote taxa it can be observed that despite not being the most common result, which is not totally surprising given the contamination event, it does come second to bacteria. Still, the great majority of the contigs identifying as Eukaryota are smaller than 1kbp for both assemblies, and with the MP producing more contigs which identify with this taxon.

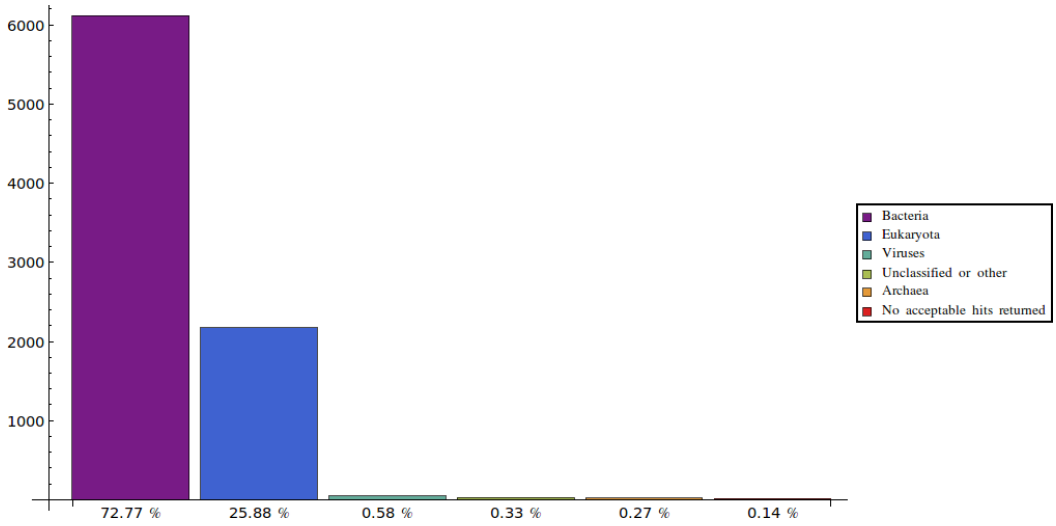


Figure 6.9: Histogram of the number of BLAST first hits for each contig by domain or match description for Mathematica Pipeline's assembly

Table 6.3: Number of BLAST hits by domain or match description for the Newbler and Mathematica Pipeline assemblies

	Number of hits	
Domain	Newbler (1)	Mathematica Pipeline
Bacteria	7991	6117
Eukaryota	1487	2176
No acceptable hits	80	12
Viruses	23	49
Unclassified	7	9
Archaea	3	23
other	18	19
Total	9609	8405

To obtain a better idea of how well each contig's best hit aligned to taxa from GenBank, both the length and fraction of each contig's matching portion were plotted (figures 6.11, and 6.12). The scatter plots in figure 6.11 show that both assemblies produced similar distribution patterns of the results for all taxa. However, there are differences worth noting, in particular for those plots corresponding to Bacteria and Eukaryota, and acccordingly these plots were redrawn to show each assembler results more clearly (figure 6.13). Comparing the Bacteria plots for both assemblies, it can be observed that the MP one has a more scattered distribution of the points. Aside from those points that are seen to overlap with Newbler ones, there is a substantial amount of points corresponding to longer contigs, whose matching portion is generally low. Regarding the Eukaryota plots, it appears that the distribution of contigs length is more homogeneous. Still, while the majority of Newbler assembled contigs were shown to match Eukaryota DNA over at least 20% of their portion, in the case of the MP, a cluster of points representing almost half of the MP contigs were observed to match below this proportion, many of which correspond to contigs larger than 1kbp. Since Newbler did not manage to assemble many contigs above this length for Eukaryota, but did for Bacteria, and given the low

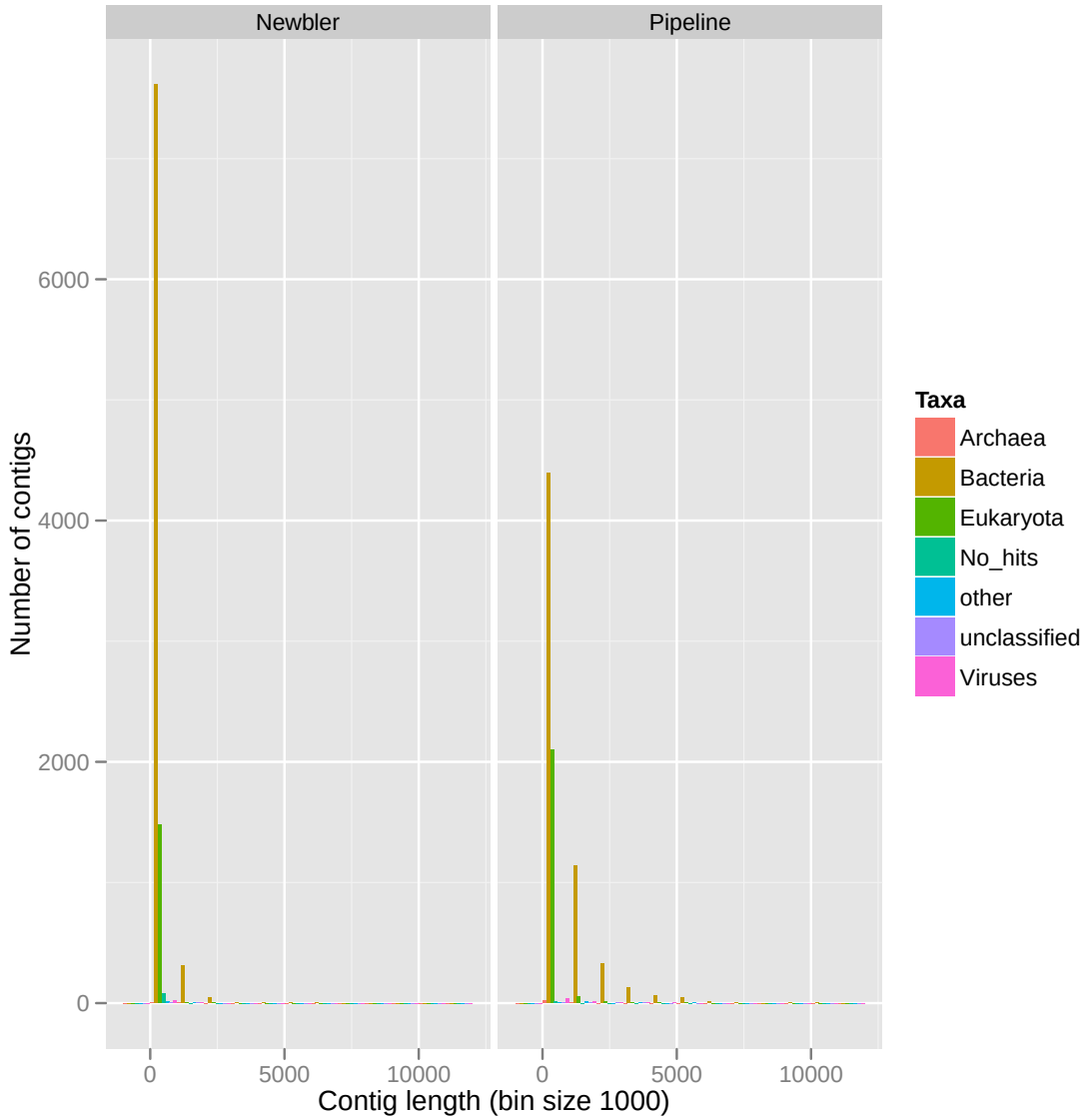


Figure 6.10: Histogram with contig length distribution by taxa for best hit contigs.

alignment portion of the contigs, which may or many not reflect the lack of suitable references in databases, one should be particularly skeptical of these set of longer contigs.

To get a more clear-cut view of each assembler's contig length distribution relative to the contig portion aligned to either Bacteria or Eukaryota taxa, two histograms were produced for each taxon (6.13). In the histogram relative to the Bacteria it is possible to observe that Newbler's contigs follow a bimodal distribution, with the modes lying on both extremes of the histogram, which account for a large amount of contigs. Specifically, more than half of the contigs are found to match over 80% of their length, while less than one third match below 20%. On the other hand, MP contigs show only a clear peak at the rightmost position of the histogram, with one third of the contigs matching above 80% of their length, and otherwise displaying a weak growing trend across the x axis, which is relatively uniform. In the histogram displaying the results for Eukaryota, Newbler contigs mostly

fall the rightmost side, where the mode is present, with two thirds of the contigs observed to lie on the 80% match region, while almost close to none in below 20% region. In the case of the MP contigs there is bimodal distribution, with most contigs laying on the leftmost side of the histogram.

Taking together, the results from the two previously succintly described figures (fig. 6.12 and 6.13), it is finally possible to shine some light on the behaviour of the two assemblies, and how exactly the two compare. To do so, it is useful to understand why the interpretation of the results, depending on the taxon being analysed, not only may, but should differ.The reason as to why this should happen, is related to the fact that a contamination event has happened, and that it should be possible to take advantage of this fact, if only a genome reference is available. In this case, the undesirable contamination event can be used as an internal control of the assembly itself.

It is precisely under this perspective that it makes sense to differentiate between the results obtained for contigs matching Bacteria, since the main source of contamination was clearly identified, and its genome is available in genomic databases, and Eukaryota, given that in the event that the lacertid sex microchromosome is indeed present, it is not guaranteed that it will share homology with other sequences available in genomic databases. In this sense, while in the case of the former taxon one might expect that contigs assembled ought to match, if well assembled, either over most of their portion, if they truly are from bacterial source, or none, if they are not, regarding the latter taxon it is truly unpredictable how well the contigs should match to GenBank taxa. Since it is not known if the lacertid shares any homology, or even to what extent, with other eukaryotes, all matches, or lack thereof, are susceptible to be interpreted as theoretically acceptable results. With this in mind, the results obtained from contigs matching Bacteria, appear to suggest that while the contigs metrics are better for the MP assembly, which produced longer contigs and has a larger N50, it would seem that it is in reality Newbler which performed the better assembly. This conclusion stems from the greater proportion of Newbler contigs which match over most of their range, as well as the reduced number of those matching over more than 20% and less than 80%, i.e. those matching over an intermediate extent. The moderate number of contigs with very low match portion however, may either represent non-bacterial contigs which matched only by chance, or it would be possible that it resulted from the assembly of reads from different taxa, or the union of two distinct repetitive regions of the same genome, i.e. chimeras. In contrast, while the MP does produce a peak for contigs matching over most of their range, there are many contigs which match over an intermediate spectre of their length, which could suggest that misassembly was involved, resulting in the production of a large amount of chimeras.

While in the case of plots for Bacteria it is possible to infer something, for Eukaryota the results are not so open to bold interpretations, mainly due to the above mentioned reasons. It would seem however, that the larger amount of contigs assembled by Newbler relatively to the MP, which show almost perfect match to Eukaryota taxa, together with the previous results that indicate that Newbler performed the bacterial assembly better, may lead one to think that in general Newbler behaves better. However, only with lab validation is it possible to understand if those long contigs with inferior levels of coverage are the outcome of misassembly, or a consequence of the lack of suitable genomes in databases for comparison.

From these results it is specially noteworthy, considering the lack of any close genome references for the lizard's sexual chromosome, the unexpected large amount of contigs identified as Eukaryota

which match over great part of their length to this taxa. In this sense, provided that they are long enough, and if shown to be conserved across several Eukaryota taxa, these could be suitable candidates for further lab validation. Furthermore, the choice of a candidate should also take into account a thorough evaluation of which assembler perform better, to avoid selecting contigs which might be chimeras.

The depth coverage per contig was also plotted for both methods (figure 6.14 for MP, 6.15 for Newbler, and a boxplot for both 8.3 in Supplementary Images). As expected, and because the MP incorporated more reads in its assembly initial dataset, the MP's depth of coverage is on average higher for all taxa. Interestingly, the contigs with no hits have a high level of coverage compared to other taxa, such as Archaea and Viruses which are similarly represented by a few amount of contigs. From the scatter plots in 6.14 it would seem however, that this coverage is largely attributable to three contigs, which have more than 200 of depth of coverage (DOC), while the majority have lower depth of coverage values.

By looking at both assembly's depth of coverage scatter plots and contrasting them, it is noteworthy to see that in the Eukaryota pane from Newbler's scatterplot, most contigs have generally depth of coverage lower than 100 DOC, while in the corresponding pane on the MP scatterplot, this value ascends to up to 400 DOC. Additionally, it can be observed that for the MP, both in the Bacteria and Eukaryota panes, the distribution of values have similar patterns and depth of coverage ranges, contrary to what is seen on the corresponding Newbler scatterplots 6.15, which show distinct patterns for both taxa. Given the presence of such striking difference of distribution patterns, one possibility to explain these results would be to assume that the MP contigs which mainly lie in the range from 100 DOC to 400 DOC in MP's Eukaryota pane, may correspond to bacterial chimeras. These would have incorporated several reads from the bacterial source, which would explain why the pattern of the MP Eukaryote pane is so similar to that observed in the Bacteria pane. However, it is also possible to entertain the possibility that the reads which did not make it into Newbler's initial reads data set, but did in the MP, could have contributed to this striking difference between assemblies. Furthermore since the portion of contigs identified as Eukaryota is particularly large, this could in some way reflect the fact that many of those excluded reads, overall shorter than the average, belonged to the lacertid. This particular association between small read size and a particular taxon, could have resulted from the use of a cell lysis step in the whole-genome amplification, which might have severely degraded, and done more warm to the naked W sex-chromosomes, than to the better protected bacteria. Moreover, given that the lacertid sex-michromosome might be full of repeats, the availability of short reads belonging to it could result in an assembly where many reads would overlap, if only by chance, further increasing the overall contig's depth of coverage for that particular taxon.

In order to see how well the contigs from both assemblies match, these were aligned to each other in pairwise fashion using BLAST. The preference of a local alignment algorithm instead of a global one was justified by the fact that the misassembly of contigs could have produced contigs which joined originally unrelated reads. A local algorithm will then be able to detect smaller but true similarities between contigs. A total of 4393 (89%) comparisons between contigs have the same BLAST result (table 6.4), the second most common combination being contigs whose best hits were identified as Bacteria in Newbler, and Eukaryota in the MP.

Since previous empirical results seemed to imply that Newbler might be better than the MP at producing well-assembled contigs, and avoiding the assembly of chimeras, it seems reasonable to expect that the alignment of contigs from both assemblies might result in a many-to-few relationship. In this type of scenario it would be often be expected for more than one Newbler contig to align to a unique chimeric MP contig, by virtue of the latter containing reads which belong to different genomic regions or taxa. However, the pattern found was the exact opposite, with a total of 4922 contigs from the Newbler assembly aligning over different lengths to 5789 contigs from the MP. This contradictory result might suggest that there are cases where Newbler does not perform well, and probably could be producing chimeras to which several MP contigs would align.

An additional test one might want to do, to infer both the quality of both assemblies and the extent of chimeras production, is to use the identification provided by each contig's best hit to GenBank and see how these agree, or disagree between contigs aligned. For example, assuming that the contigs from Newbler are not chimeras, it is expectable that when aligning the MP contigs to Newbler's, there should be a very low chance of MP contigs identified as belonging to different taxa from Newbler contig, to match over the portion identified as so. In addition, if two MP contigs do agree between their assigned GenBank taxon id, and align to the same contig, it would seem reasonable that, if the assembly went well, these should have been merged in the assembly. From looking at the contigs that aligned between assemblies, it was possible to identify a total of 218 Newbler contigs out of the 4922 Newbler contigs, which were overlapped by at least two different taxa, where one of the contigs was assigned as belonging to a different taxon from the Newbler contig. In contrast, the number of MP contigs found to be is much lower, with only 93 contigs out of the 5789 aligned. This again, as in the previous case, reinforces the idea that in some ways the MP may have performed better than Newbler. By analysing if at least two contigs of one assembly mapped to the contig of another, all sharing the same taxon, it was observed that both the assemblies obtained similar numbers. Specifically, a total of 2034 MP contigs aligned each to at least two Newbler contigs, while 2197 Newbler contigs aligned each to at least two MP contigs. Thus, if the contigs, to which more than one contig is being aligned, are not a chimera either of two similar repeats from different parts of the genome, or produced by combining the products of more than one taxon, it seems that to some extent both assemblies were equally ineffective in the assembly step.

The best hit BLAST results obtained from each assembly's contigs that could be aligned to each other were plotted to see how well they would agree. The length of either assembly's smallest contig and its portion matching the corresponding contig in the opposite assembly, were plotted for each comparison (figure 6.16). From the scatterplots in figure 6.16, it can be observed that some of the contigs which did not get any hits on GenBank, are now seen to match to contigs identified as Bacteria and Eukaryota. It is also interesting to note that some of longer contigs identified as Eukaryota in the MP, and which had lengths above 1kbp, do overlap, even if partially with some of Newbler's contigs identified as Bacteria, although it is not clear if both contigs, or only one, are chimeras.

Given that the identity of the contigs aligned between assemblies is obtained by their best hit on GenBank, and thus may only correspond to part of its length, which is not necessarily the one aligned between assemblies, an extra set of scatter plots was generated. In these the contigs shown have to match at least 90% of their width to their best GenBank hit (figure 6.17). As expected, by increasing the stringency of the identity, much of the disagreements stemming from the chimeras disappeared, revealing that possibly a lot of contigs, for which there is a higher confidence of having been well

assembled given their identity scores, could be overlapped and merged.

Table 6.4: Comparing top BLAST hits for contig pairs that align between the two assembler outputs.

BLAST results (Newbler/MP)	Number of contigs
Archaea/Eukaryota	2
Bacteria/Archaea	5
Bacteria/Bacteria	3901
Bacteria/Eukaryota	390
Bacteria/No hits	2
Bacteria/other	6
Bacteria/unclassified	9
Bacteria/Viruses	12
Eukaryota/Bacteria	48
Eukaryota/Eukaryota	492
Eukaryota/other	2
Eukaryota/Viruses	1
No hits/Bacteria	7
No hits/Eukaryota	13
Other/Bacteria	1
Other/Other	9
Unclassified/Bacteria	1
Unclassified/Eukaryota	1
Unclassified/Unclassified	4
Viruses/Bacteria	2
Viruses/Eukaryota	4
Viruses/Viruses	10

Mapping to the bacteria genome

To further validate the BLAST results, and try to shed some light on the contigs' assembly status and differences between the two assemblies, the BWA mapper (Li and Durbin, 2009) was used to map the contigs of both assemblies to the *S. maltophilia* bacterium genome. For this purpose the BWA-SW algorithm, particularly suited for longer reads, was used.

The percentage of the 4,851,126 bp bacterium genome positions covered at least once in each assembly are displayed, for both BLAST and BWA alignments in table 6.5. It should be noted that the extremely repetitive nature of the bacterium genome may underestimate and/or overestimate these numbers for two reasons. Firstly, as described earlier, the presence of repeats throughout the genome, increases the chances of incorrectly assembling into a single contig, reads sharing the same repeat motif, but which come from different genomic regions. This possibly inhibits the mapping of the contig, which leads to an underestimation of the real breadth of genome coverage. Conversely, the contigs which were generated from regions contained within repeat motifs will often lack the resolution

to be correctly mapped to only one place, whence they truly originate from. In this instance, if a region which was not sequenced, but is found to be an exact copy of another region of the genome, or otherwise extremely similar, which was sequenced, the former may appear to have been sequenced, unless the depth of coverage tells another story, creating the perception that a larger part of the genome was covered.

Table 6.5: Percentage of breadth of coverage attained by the contigs mapped to *S. maltophilia*, by assembler and alignment tool.

	Mathematica Pipeline	Newbler
BWA	23.6%	27.4%
BLAST	25.3%	28.6%

The comparison between BLAST and BWA results showed a high level of concordance among the contigs mapping to *S. maltophilia* in both assemblies, with only a small amount of contigs being unique to each method (table 6.6).

Table 6.6: Number of contigs mapping to *S. maltophilia* by alignment tool and assembler, and the percentage of contigs uniquely discovered by each tool by assembler

	Mathematica Pipeline (% Unique)	Newbler (% Unique)
BWA	3783 (5.02%)	3603 (6.71%)
BLAST	3671 (2.17%)	3392 (0.91%)

Furthermore, the two alignment tools also show consistency between the proportion and length of each contig's best hit by assembly, (fig. 6.18 and fig. 6.19) with correlation coefficients r higher than 0.97 and with a p-value of $2.2*10^{-6}$.

To obtain an idea of how much more breadth wise the genome could have been covered, the initial set of processed reads input to each assembly (instead of their output contigs) was also mapped with BWA to the bacterium genome. Table 6.7 shows that the amount of positions covered by both sets of processed reads is similar, with the processed reads used as the input for the MP showing a little more coverage than Newbler's input dataset. This is not surprising given that the MP managed to incorporate a larger number of reads as its entry dataset. More importantly, these results show that the reads were able to cover an extra $\sim$20% of the bacterium genome over the mapping performed with contigs. This can further be observed in the representation of the bacteria genome obtained using Circos (Krzywinski et al., 2009), where both the reads, and contigs, for each assembly, which had been shown to map to it, were plotted. In these bacterial genome representations it is possible to observe a higher degree of patchiness for contigs compared to the one relative to the reads, which agrees with the idea that the breadth of coverage from the former is not as good as the latter (figure 6.20 and figure 6.21).

Although some of this difference may be accounted by a higher aptness of shorter sequence fragments to map to the genome, the difference of coverage observed seems too high to be dismissed. This remains true even if some reads map by chance, or because their shorter size allows them to deal better with the variation there may be between the reference bacterium genome, and the bacteria whence the reads originate. Such an observable difference could then be interpreted as a measure

Table 6.7: Reads and contigs *S. maltophilia* breadth of coverage by assembly mapped with BWA

	Mathematica Pipeline	Newbler
Contigs	23.6%	27.4%
Reads	44.0%	41.1%

of the inefficiency of both assemblers to perform the best assembly possible. Possible explanations for this are the prematurate incorporation of the reads into a contig, inhibiting their further use in other contigs, reads misassembly, algorithm greediness, or lack of depth of coverage to support the assembly of the correct set of reads.

Besides showing a more fragmented coverage of the contigs relative to reads, the Circos bacterial genome representations contain red lines providing visual information about the contigs mapping to more than one place in the genome. This corresponds to the several lines in the center of the each circle connecting different genomic regions. The denser array of connections displayed by the MP assembly suggests that either the contigs suffer from misincorporation of unrelated reads, so that the contigs will map to several places, or it can reflect the real repetitive nature of the bacterial genome, whose motifs should have been more often captured by individual MP contigs given their larger size. Indeed, it is particularly believable, that in the MP, the greediness of the algorithm probably makes it more prone to be expose to the mentioned phenomenas. This would result in the assembly of chimeras, particularly out of reads which are found to share repetitive regions present throughout the genome. If the contigs of the MP are then chosen for further lab validation, one should be decidedly wary of this fact, or otherwise risk designing primers for a contig which represents not one, but two or even several distinct genomic regions, thereby increasing the chances of failure in the following validation step. In this regard, at least making sure that the contigs possess a depth of coverage relatively similar over their width, and which resembles the background level, or in the case here is homology with other taxa, that the primed region is found to match to a relevant taxon in its entirety, might increase the chances of success.

In both mapping methods it was noteworthy the extent of soft-masking, which is just a process by which certain parts of the reads are masked, so essentially ignored, during the mapping so that it can take place, that the edges of the reads were subjected to. This might be evidence of misassembly which can result, as previously noted 1.2, from the incorporation into the same contig of the different unique nucleotidic regions neighbouring repeat motifs from different genomic regions. Alternatively, it could support the idea that there is an increasing chance of misassembly as the extension proceeds in both directions and moves further away from the initial and core part of the contig, which is usually well supported by a greater amount of reads.

The MP appeared to be particularly affected by this problem, presenting a higher amount of soft-masked base pairs compared to Newbler (table 6.8 and table 6.9).

Table 6.8: Statistics for the soft clipped regions that flank the contigs for both assemblies

Side	Metric	Pipeline	Newbler
Left	Mean	211.3	33.6
	SD	492.2	157.0
	Max	7126	4451
Right	Mean	235.7	36.0
	SD	563.3	159.6
	Max	8188	3908
Both	Mean	223.5	34.8
	SD	529.0	158.3
	Max	8188	4451

Table 6.9: Number of base pairs soft clipped in contigs produced by both assemblies

	All contigs	Only best hits
Newbler	419363 (3.4%)	169618 (1.5%)
Mathematica Pipeline	1467881 (30.0%)	419363 (8.6%)

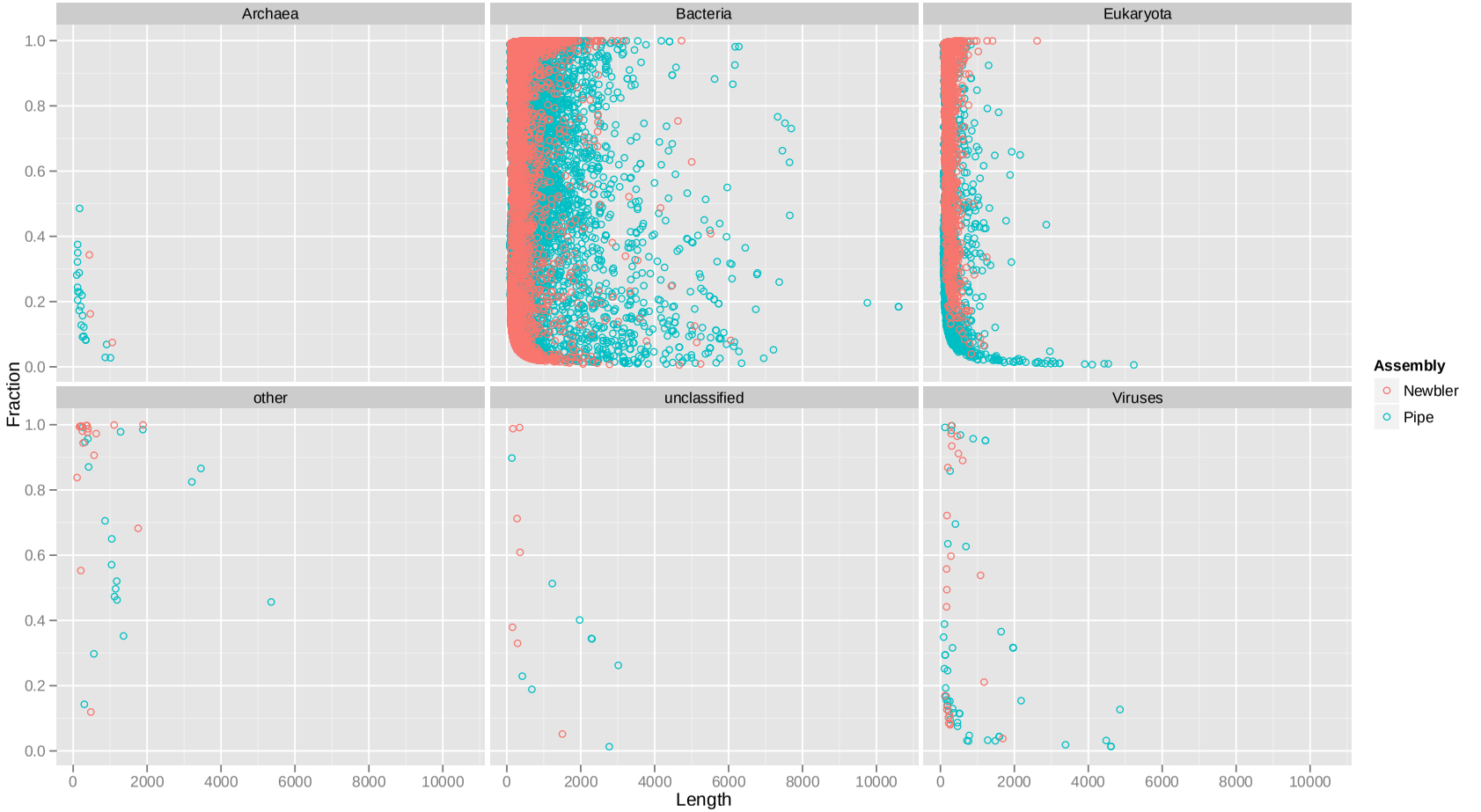


Figure 6.11: Scatter plot showing for both assemblies the length distribution of the best hit contigs and the portion matched by the respective taxa.

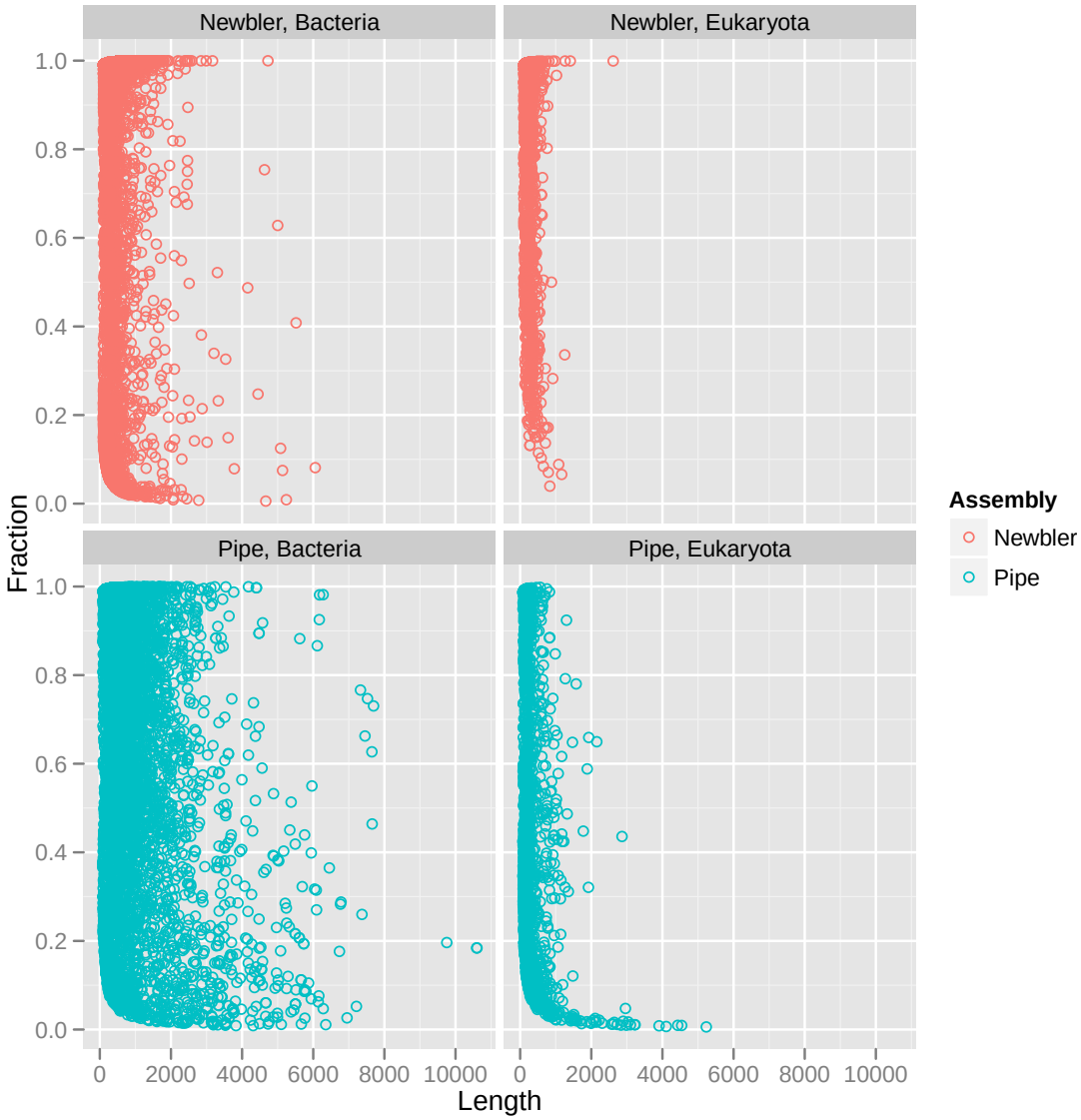


Figure 6.12: Scatter plot showing for both assemblies the length distribution of the best hit contigs only to bacteria and Eukaryota, and the portion matched by the respective taxa.

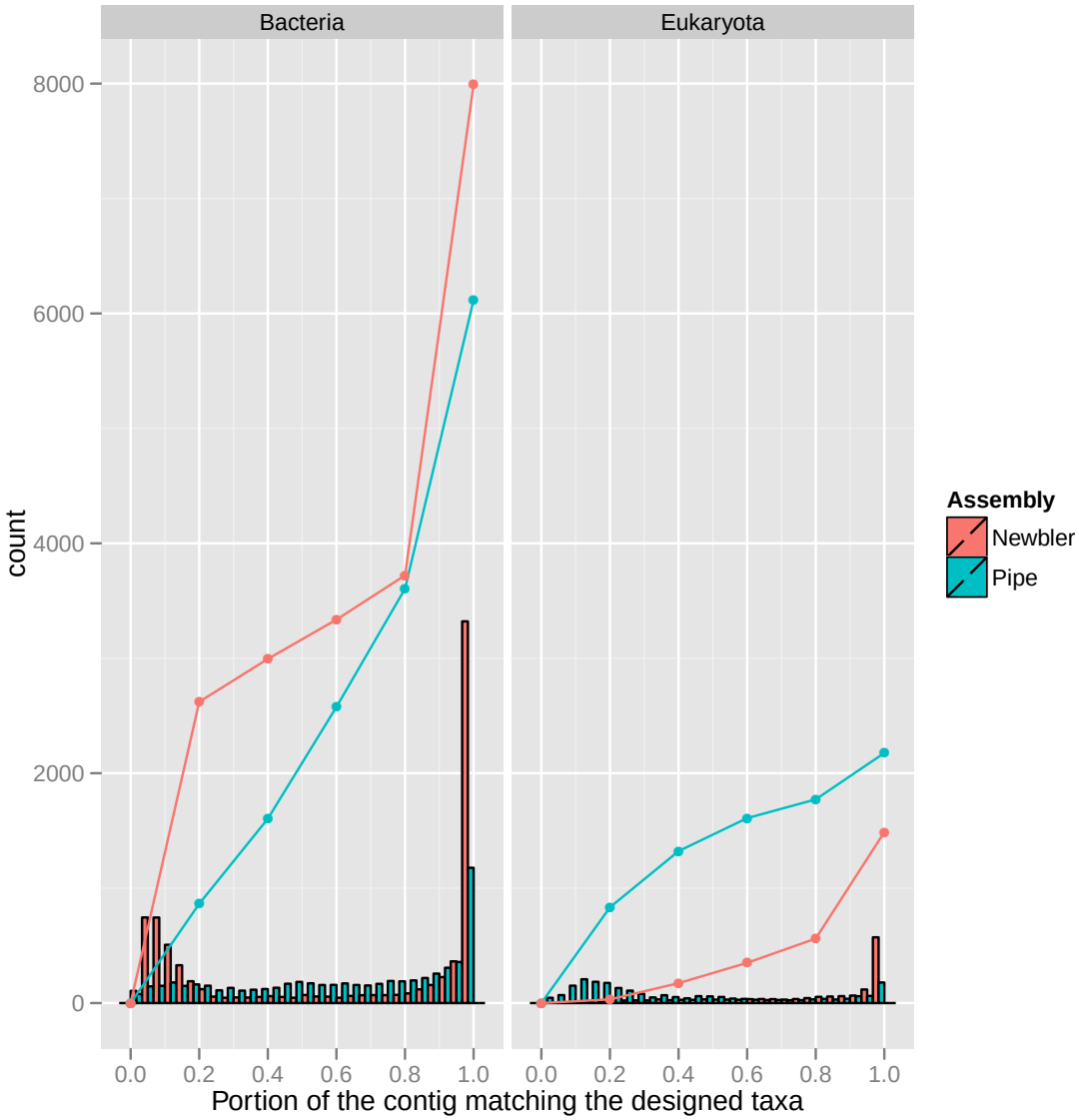


Figure 6.13: Histogram displaying number of contigs by the portion their best hit matches to Bacteria and Eukaryota. The lines represent the cumulative number of contigs.

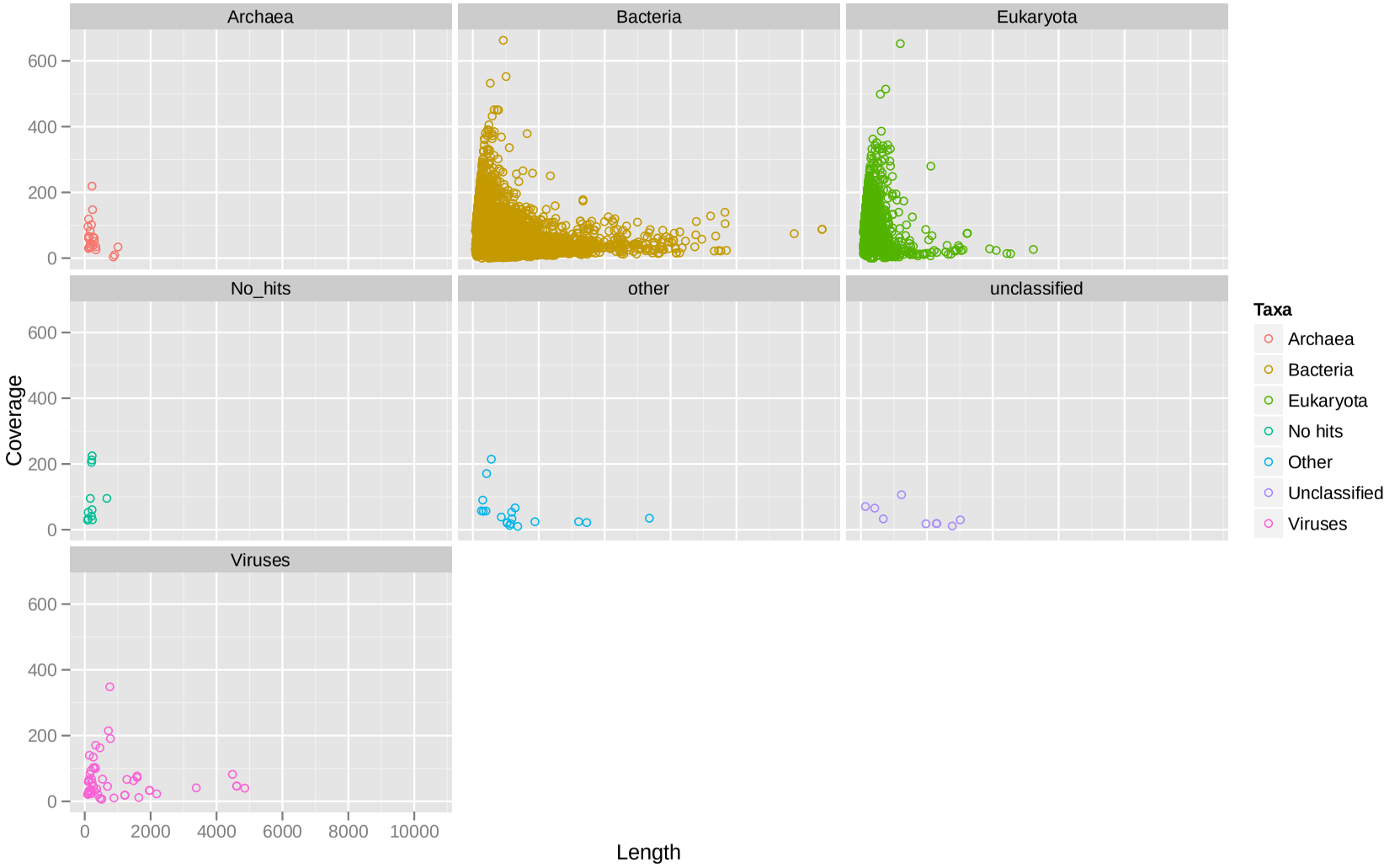


Figure 6.14: Scatter plots showing depth of coverage distribution (y axis) versus contig length (x axis) by taxa for the Mathematica Pipeline.

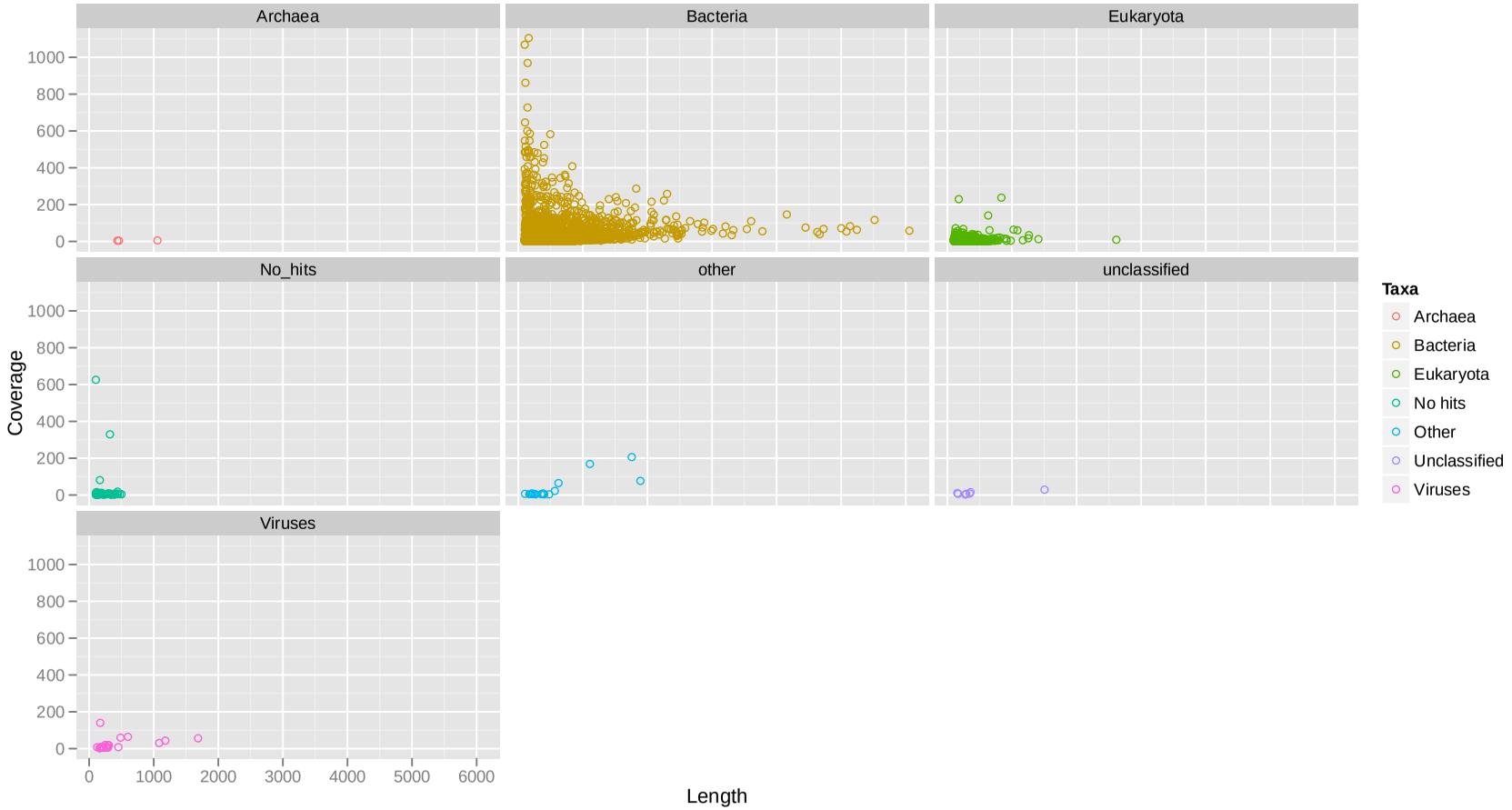


Figure 6.15: Scatter plots showing depth of coverage distribution (y axis) versus contig length (x axis) by taxa for Newbler.

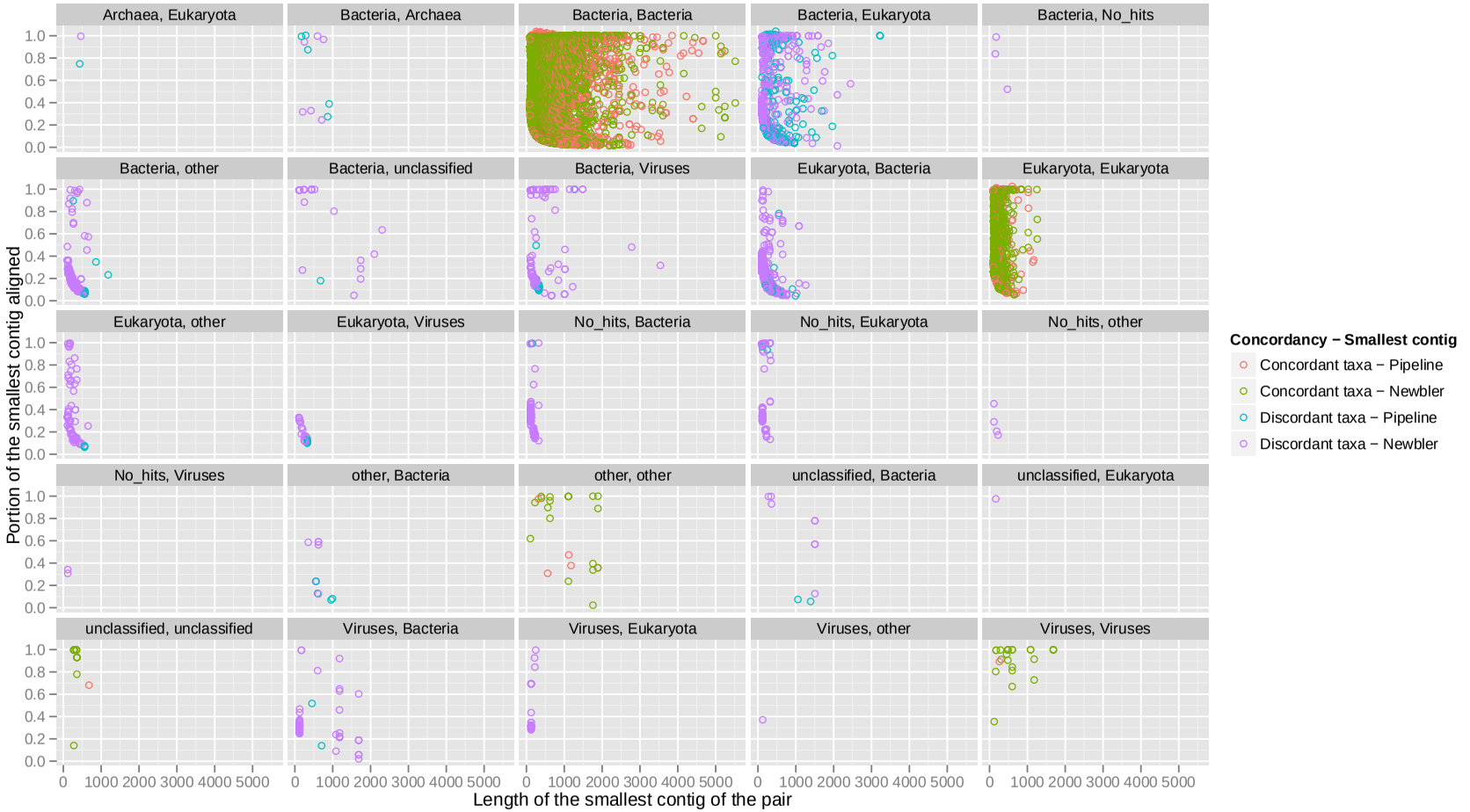


Figure 6.16: **Scatter plots displaying all possible combinations of the top BLAST result obtained by contigs of both assemblies which could be aligned in pairwise fashion.** Each pane corresponds to a combination of the top BLAST result from one contig from Newbler, and one contig from Mathematica Pipeline, which proved to align the best from all the contigs from each assembly. Plots with pink and green dots display those contigs for which the taxa is concordant between the two assemblies. Plots with blue and purple dots represent those contigs for which the taxa is discordant between the two assemblies. Each dot represents the length (x axis) of the smallest of the two contigs from both assemblies, and the portion of the smallest contig of the pair aligned between the two contigs (y axis). Pink and blue when Newbler has the smallest contigs of the pair, purple and green otherwise.

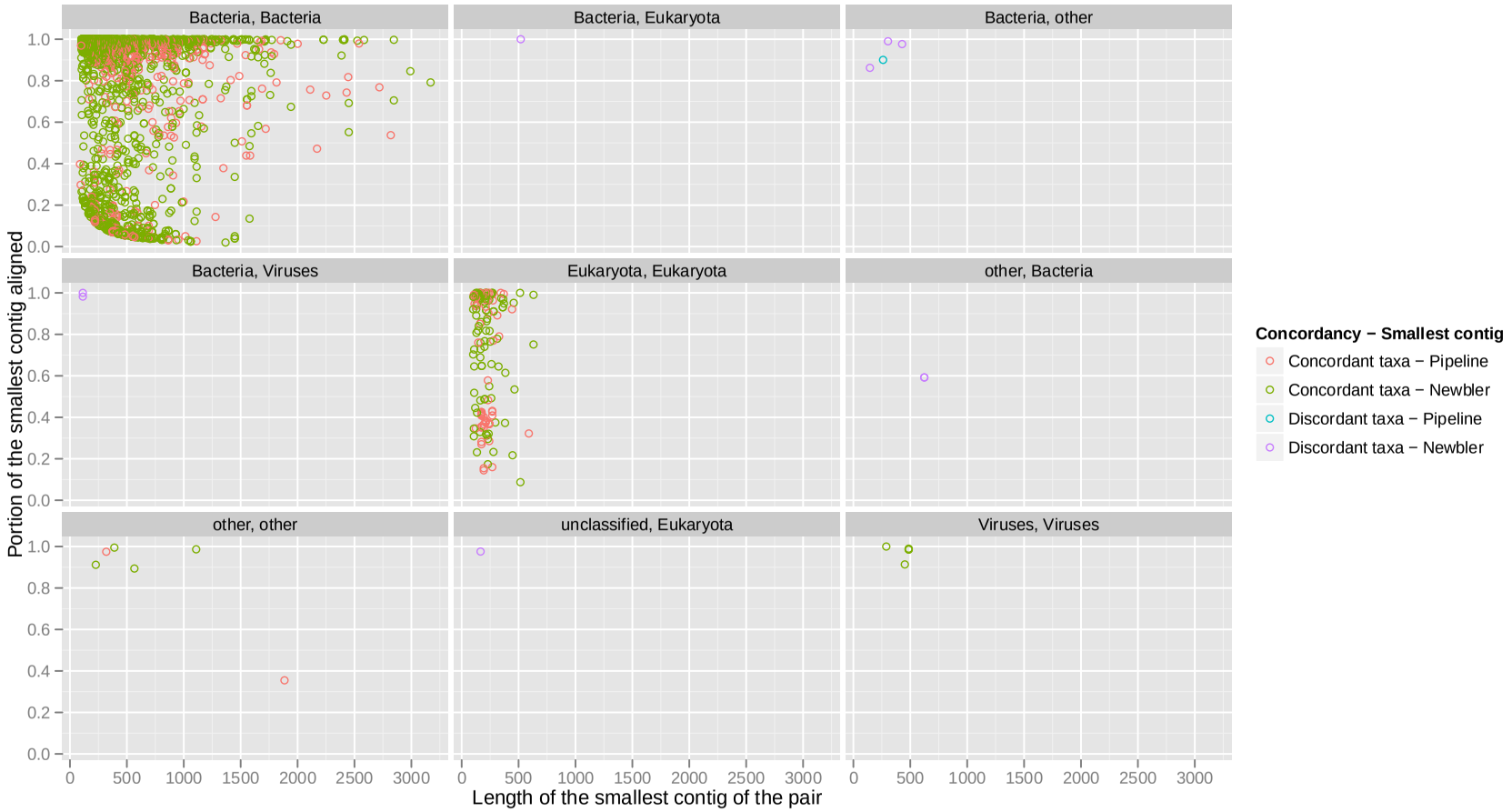


Figure 6.17: **Scatter plots displaying all possible combinations of BLAST results between the two best matching contigs from each assembly, which had aligned over at least 90% of their width to its assigned taxon.** Plots with pink and green dots display those contigs for which the taxa is concordant between the two assemblies. Plots with blue and purple dots represent those contigs for which the taxa is discordant between the two assemblies. Each dot represents the length (x axis) of the smallest of the two contigs from both assemblies, and the portion of the smallest contig of the pair aligned between the two contigs (y axis). Pink and blue when Newbler has the smallest contigs of the pair, purple and green otherwise.

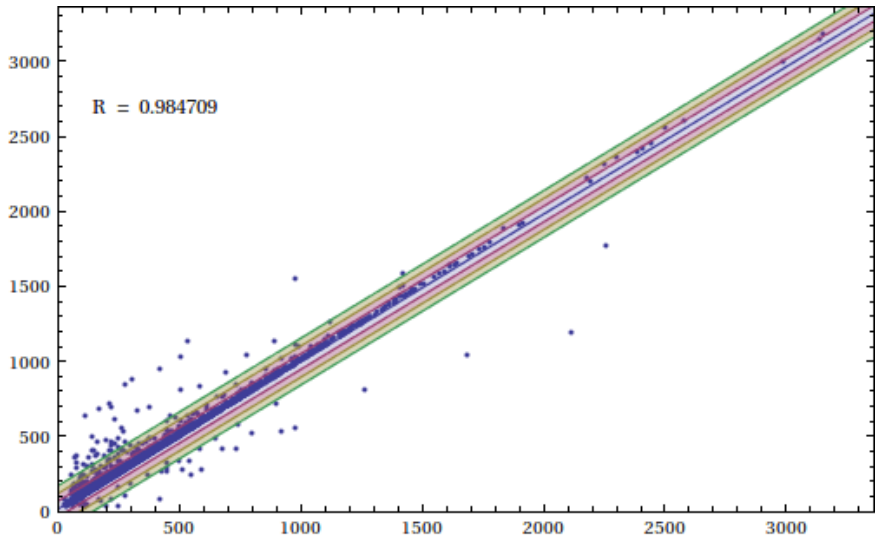


Figure 6.18: Newbler's correlation of contig size from best hits obtained by BWA and BLAST. Each differently transparent coloured region represents a further standard deviation away from the last region or the original line.

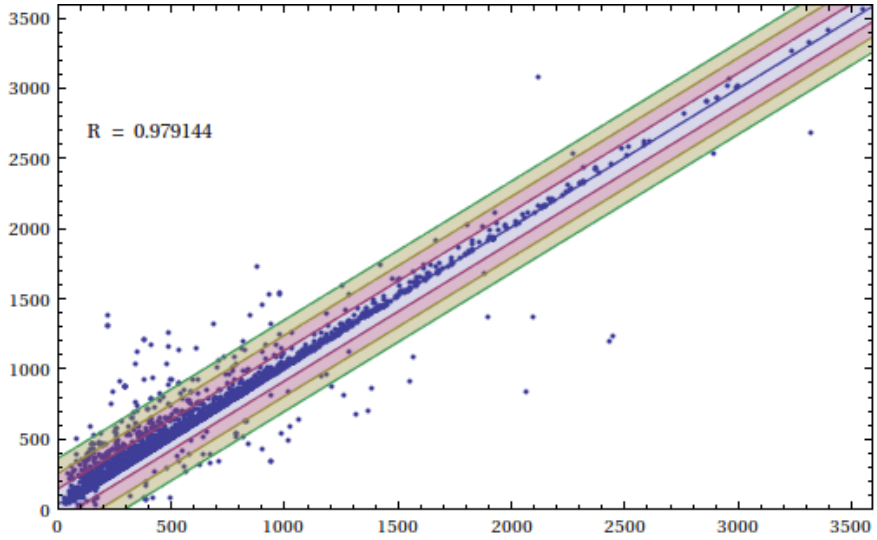


Figure 6.19: Mathematica Pipeline's correlation of contig size from best hits obtained by BWA and BLAST. Each differently transparent coloured region represents a further standard deviation away from the last region or the original line.

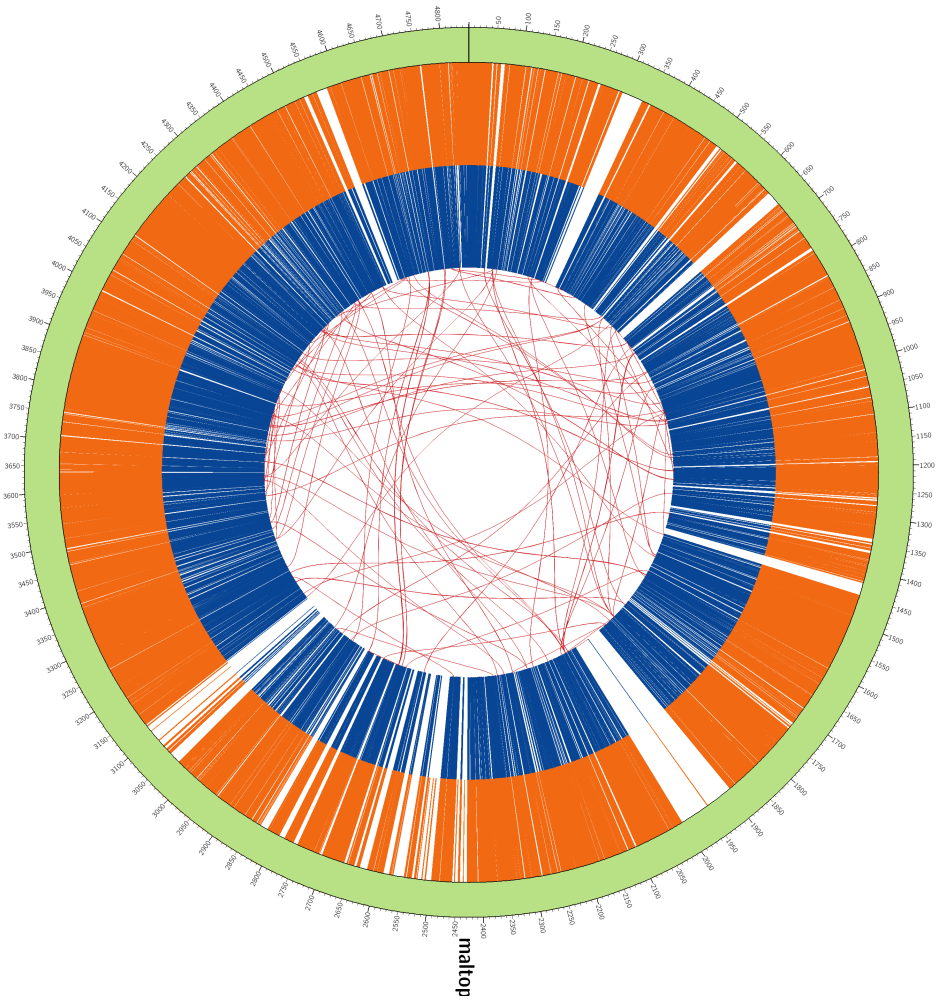


Figure 6.20: Representation of *S. maltophilia*'s reference genome, with contigs (dark blue) and reads (orange), respectively generated and used by Newbler, plotted. The red lines in the center of the figure connect positions to which the same contig mapped.

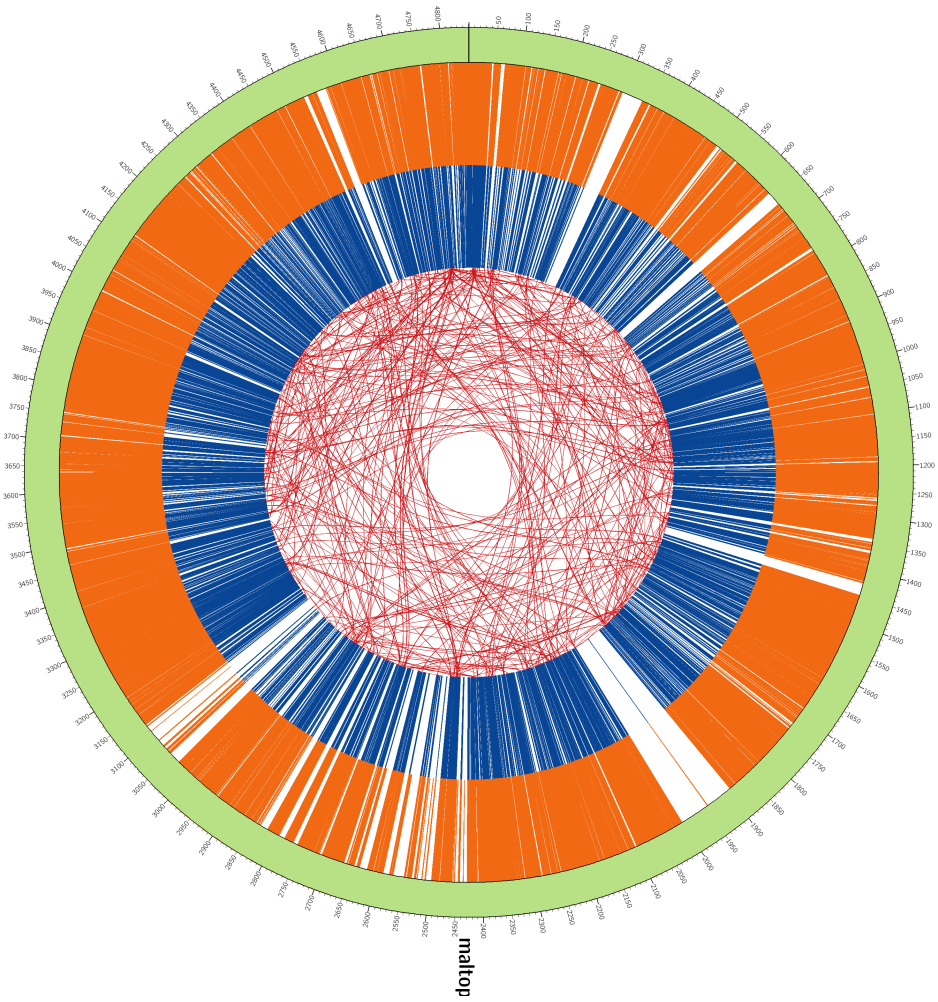


Figure 6.21: Representation of *S. maltophilia*'s reference genome, with contigs (dark blue) and reads (orange), respectively generated and used by MP, plotted. The red lines in the center of the figure connect positions to which the same contig mapped.

Chapter 7

Laboratory validation

The contigs produced by both assemblers proved to be identified mostly as bacteria, with *Stenotrophomonas maltophilia* making up the most part of it, as indicated by the overwhelming amount of contigs identified primarily as so in the BLAST results. Nonetheless, several other contigs were found that did not map to bacteria but instead to Eukaryote taxa, and could belong to the lizard's W sex microchromosome.

Given the lack of close and reliable genome references to ascertain the validity of this hypothesis, lab methods were employed to try and validate this hypothesis.

The method appointed was PCR, which allows for cheap and reasonably fast multiple testing of molecular sequences. The contigs assembled, and not identified as being bacteria, were used as template for primer design, in order to ascertain if they amplify samples from the same species as the lizard which had theoretically been sequenced.

To perform this assessment, a total of 19 contigs were selected. These were the longest contigs shown to match exclusively to Eukaryota taxa, and whose BLAST results had sex related keywords. The keywords encompassed names of genes involved in sexual pathways, sequence elements known to be highly conserved in Eukaryote taxa and related to sex such as HMG-box, as well as any other terms which could in some way be related to sex. Furthermore, at least for one of these 19 contigs, the target region for amplification was made sure to be covered by unique reads in the assembly, in order to infer if eventual amplifications problems could be linked to the assembly step. In addition, one contig of *Stenotrophomonas maltophilia* was included as a control.

In order to design the primers a stand-alone version of primer3 (Rozen and Skaletsky, 1998), a tool specifically created for primer design, was used to generate a total of 20 forward and reverse primers out of each contig.

To test the primers, it was decided that tail samples from a male and a female *Eremias velox* lizard, the female having been previously the source for the microdissected tissue, would be selected in addition to tail samples from other male and female lizards known to possess the ZZ/ZW sex system. The former samples, if shown to amplify, could be interpreted as evidence that the contigs were well assembled, and that lizard reads were present in the sequence data. Additionally, the inclusion

of a male individual, which does not possess a W chromosome in contrast to the female, could help pinpoint the amplification specifically to the W chromosome, giving some support to the idea that the W sex microchromosome was indeed sequenced and well-assembled, in the case that only the female sample is found to be amplified. The non-*Eremias* samples should allow us to look for homology among different lizard taxa, and in the case only the female samples are amplified, to infer that the regions amplified are restricted, and shared among, W sex microchromosomes.

For this purpose tail tissue samples from a male and a female *Darevskia valentini* lizard as well as from a *Darevskia raddei* parthenogenic female lizard were selected for DNA extraction in addition to the two *Eremias velox* samples, from each one of the two individuals, male and female. The extraction of DNA from both *Eremias velox* samples was replicated, in order to account for possible issues in the extraction step, and serve as double positive control in the case the lacertid was found to be amplified. All samples were extracted using the saline methods (Sambrook et al., 1989).

As a mean of experimental control, a pair of primers targeting a conserved region of the MC1R gene (Pinho et al., 2010), was also used to amplify the seven samples in every PCR. These were included with the intention of ensuring that the PCR was performing well, since they had been shown to amplify the seven samples. In this way it should be possible to more promptly detect problems with the reagents.

Since the primers utilized on the course of this experiment were designed *de novo*, optimal PCR amplification conditions such as sample and reagent amount, or temperature were unknown. As such, the conditions and reagent concentrations observed to previously work with the positive control primers were initially adopted.

The reagents initially used and their concentrations are displayed in table 7.1. This standard PCR protocol consists of an initial 3 minute cycle at 93° C, followed by 20 cycles of temperature touchdown, each comprised by a 30 second denaturation period at 93° C, an annealing temperature step performed at 65° C in the first cycle, which decreases at a rate of 0.5° C per cycle, and an extension step at 72° C for one minute. Afterwards a routine of 15 cycles is performed, where each cycle starts with 30 seconds at 93°C, followed by stable annealing at 55° C, and is left to extend for one minute. Finally, at the very end of the PCR there is a 10 minutes extension step at 72° C.

Table 7.1: Reagents and quantities initially used to perform PCR.

Reagents	Quantities (μl)
H_20	31.2
Buffer	12
2.5 mM $MgCl_2$	7.2
0.4 mM dNTPs	2.4
0.3 μM Forward primer	0.9
0.3 μM Reverse primer	0.9
1U GoTaq Taq polymerase	0.6
DNA	0.5

These conditions and the MC1R primers were first used to find the optimal sample concentration, with the following sample extractions dilutions being tested, 1:5 ,1:10, 1:20 and 1:40. All samples

were run on a 0.8% agarose gel, and for each sample the dilution shown to produce well defined and not overly saturated bands was used throughout the remainder of the experimental course. This PCR protocol was then used to test the 7 samples with each of the 20 primer sets, plus the MC1R primers.

The first batch of PCR trials with the twenty designed primer pairs failed to amplify all samples, with exception of the primers designed from bacteria which were seemingly able to amplify the female *Darevskia* samples. On the other hand, both the negative and positive controls performed properly (i.e. no amplification occurred in the negative control wells, and amplification occurred in all positive control wells with a sample).

The unexpected amplification of female *Darevskia* samples with primers designed from bacteria, and the absence of amplification in all other cases, aside from the positive control, lead to a new round of PCRs which also failed to amplify the samples.

Since in part this could be due to the PCR conditions not being optimized for the primers used on the course of the experimental work, further testing was done to exclude this hypothesis as a possible explanation for the absence of amplification.

A more comprehensive array of tests, using the the designed primers, was then performed with a gradient of five annealing temperatures (56° C, 58° C, 60° C, 62.4° C, 64.3° C, the well at 56.7° C was always used for the gradient negative control), and three $MgCl_2$ concentration gradients (0.6, 1.2 and 2.4, half, respectively half, the same, and double of the previously used amount). The sample set was reduced to one of the *Eremias velox* female samples, to be tested with all the primers, and the *Darevskia raddei* sample, which was previously observed to produce the best amplification results with the positive control primers.

All PCRs failed to amplify with primers designed from contigs presumed to belong to Eukaryota taxa, while both PCR runs with $MgCl_2$ of 1.2 and 2.4 μl, and annealing temperatures of 56° C and 58° C proved to work with the primers designed for bacteria. These results were further validated by using the latter set of primers to try and amplify the other female Eremias sample and achieving the same results.

In order to clarify why the primers designed from a bacterial contig were able to amplify DNA from several samples, two samples of amplified DNA from *Darevskia raddei* were sequenced with the Sanger method, using these same PCR in the reaction.

For each sample, the reverse and forward primers were used separately, resulting in a total of 4 sequenced fragments, two for each sample, with sizes ranging between 657 and 709 bp. After BLASTing the sequenced fragments, these were shown to have several hits with high degree identity against a large range of bacteria, including *Stenotrophomonas maltophilia*. However, the portion of the best match among all the sequenced fragments to *Stenotrophomonas maltophilia*, was found to be 399 bp, with an identity of ∼83%, both inferior to the metrics presented by other bacteria best matches, such as for example *Escherichia coli*, whose best match of the same fragment sequence, which was 661 bp, aligned over 629 bp to its genome, with an identity of ∼89%. These results suggest that, either during the DNA sample extraction, or afterwards, some bacteria taxa must have contaminated the samples, or alternatively that it was from the lizards' own microbiome. Moreover, since the best matches of the sequenced fragments to bacterial genomes, was found to be to others that

not *Stenotrophomonas maltophilia*, it is probable that the source of PCR contamination originated from other bacterial sources, which have some degree of similarity with *Stenotrophomonas maltophilia* at the primed regions.

Chapter 8

Main conclusions and future prospects

While the combination of NGS, cytogenetic techniques and enrichment methods certainly represents a very powerful and ingenious approach to further advance the genomics field, it can also be affected by drawbacks and complications. In this trial, a putative contamination event which took place even before the sequencing, may have compromised what could otherwise turned out to be a fruitful approach, and influenced the course of the experimental work.

The lack of positive results on the lab even in the presence of a reasonable amount of BLAST results suggesting the presence of contigs matching Eukaryote taxa among the assembled contigs, warrants further data exploration and validation approaches. In particular, given that the validation portion of the trial was severely time constrained, not only was the selection of contigs *in silico* possibly not optimal, but also their *in vitro* validation was not as thorough as it could be. By addressing these issues, it is possible that the results may see some improvement. Additionally, it might be interesting to try to look for, and identify, other steps of the trial which could be further improved.

Future work will then be based on two fronts: computational and experimental. On the computational side, the assemblers should be compared with genomes of different complexities, in order to see how well they perform and in which conditions. It should be particularly interesting to see how differently they perform with Eukaryote taxa simulated genomes since this group of organisms should pose a different array of problems to the assembly compared to bacterial genomes, and could in some way rebut the hypothetical advantage of either assembler for these cases. Furthermore, since the empirical results do suggest that chimeric contigs might be indeed a prevalent problem, more thought and effort should be put in their identification. Addressing this issue would probably involve implementing extra measures to take better advantage of the depth of coverage information, and check that the depth of coverage is relatively constant throughout the contig, in addition to increase the stringency on the contig assembly step, byonly admitting only reads that map uniquely to a contig. Additionally, more stringent filters must be applied to the selection of the contigs to be validated in the lab, namely by attributing some sort of a score reflecting the confidence of each contigs' assembly. Lastly, an effort should also be made to include more contigs whose region to be amplified in the lab validation step,

is wholly contained by single reads, so that it is possible to discern if the problem might lie on the assembly, or on the sequence data.

On the experimental side, other methods can be employed to validate the contigs that not PCR. An example of such methods would be the design of FISH probes using the contigs as templates. This method would allow to see if the contigs have a complementary sequence in the genome of interest, and if so to pinpoint the chromosome and the relative position within it. In addition, to validate the assemblies generated, smaller regions of the W sexual microchromosome could be sequenced with the more reliable Sanger sequencing method. Complementarily, if funding is available, another sequencing effort should be taken, if possible in more controled conditions to avoid the risk of contamination. In this sense, some ponderation should also be put into to what extent the use of cell lysis in the WGA protocol might have damaged the already small chromosomes, which would explain the lower size of the reads relativel to the initial expectaction, and if more suitable WGA options should be sought.

If shown to work, in the course of future efforts, this trial should result in the development of, for example, useful sex markers specific to the W chromosome. The development of such markers could prove particularly useful if these are found to be conserved across lacertid taxa. If they are, it should be possible to sex individuals, in addition to perform comparative genomics and population genetics studies in more depth and in a larger scale.

For this reason, any steps toward the improvement of this type of approach, particularly on the presence of adversities, should be actively sought and explored. This applies specially to any negative or lack of results which should be taken as an opportunity to further learn more about problems involved with the approach, and to refine the existing methods or develop new ones. Doing so will surely set a path leading to better data production and analysis.

Additionally, in the future, this same approach could be employed not only in other, non-sex-, microchomosomes, which represent a less complex task, but also in macrochromosomes. If possible, it should also be extended to other taxa, including to model taxa for which the reference is available, effectively allowing for a more objective results validation.

Bibliography

Silvia G Acinas, Ramahi Sarma-Rupavtarm, Vanja Klepac-Ceraj, and Martin F Polz. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Applied and Environmental Microbiology*, 71(12):8966–8969, 2005. URL `http://aem.asm.org/cgi/content/abstract/71/12/8966`. 10

Arianne Y K Albert and Sarah P Otto. Sexual selection can resolve sex-linked sexual antagonism. *Science*, 310(5745):119–121, 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/16210543`. 29

Jessica Alföldi, Federica Di Palma, Manfred Grabherr, Christina Williams, Lesheng Kong, Evan Mauceli, Pamela Russell, Craig B. Lowe, Richard E. Glor, Jacob D. Jaffe, David a. Ray, Stephane Boissinot, Andrew M. Shedlock, Christopher Botka, Todd a. Castoe, John K. Colbourne, Matthew K. Fujita, Ricardo Godinez Moreno, Boudewijn F. ten Hallers, David Haussler, Andreas Heger, David Heiman, Daniel E. Janes, Jeremy Johnson, Pieter J. de Jong, Maxim Y. Koriabine, Marcia Lara, Peter a. Novick, Chris L. Organ, Sally E. Peach, Steven Poe, David D. Pollock, Kevin de Queiroz, Thomas Sanger, Steve Searle, Jeremy D. Smith, Zachary Smith, Ross Swofford, Jason Turner-Maier, Juli Wade, Sarah Young, Amonida Zadissa, Scott V. Edwards, Travis C. Glenn, Christopher J. Schneider, Jonathan B. Losos, Eric S. Lander, Matthew Breen, Chris P. Ponting, and Kerstin Lindblad-Toh. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, pages 5–9, August 2011. ISSN 0028-0836. doi: 10.1038/nature10390. URL `http://www.nature.com/doifinder/10.1038/nature10390`. 25

Can Alkan, Saba Sajjadian, and Evan E Eichler. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1):61–65, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21102452`. 7

Stephen F Altschul. BLAST Algorithm. *ENCYCLOPEDIA OF LIFE SCIENCES*, pages 1–4, 2005. doi: 10.1038/npg.els.0005253. URL `http://dx.doi.org/10.1038/npg.els.0005253`. 37

E Nicholas Arnold, Oscar Arribas, and Salvador Carranza. Systematics of the Palaearctic and Oriental lizard tribe Lacertini (Squamata: Lacertidae: Lacertinae), with descriptions of eight new genera. *Zootaxa*, 1430(1430):1–86, 2007. ISSN 11755326. URL `http://www.mapress.com/zootaxa/2007f/z01430p086f.pdf`. 26

V N Arronet. The karyotype of the lizard Ophysops elegans. *Tsitologiia*, 10(1):120–122, 1968. 27

Erik Axelsson, Matthew T Webster, Nick G C Smith, David W Burt, and Hans Ellegren. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Research*, 15(1):120–125,

2005. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=540272&tool=pmcentrez&rendertype=abstract. 26

W M Barnes. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proceedings of the National Academy of Sciences of the United States of America*, 91(6):2216–2220, 1994. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=43341&tool=pmcentrez&rendertype=abstract. 9

G P Bates, B J Wainwright, R Williamson, and S D Brown. Microdissection of and microcloning from the short arm of human chromosome 2. *Molecular and Cellular Biology*, 6(11):3826–3830, 1986. 22

Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and David L Wheeler. Genbank: update. *Nucleic Acids Research*, 32(Database issue):D23–D26, 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/14681350. 4

Stephen Bentley. Genomic 'valleys of death'., 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18340971. 2

F R Blattner. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, 1997. ISSN 00368075. doi: 10.1126/science.277.5331.1453. URL http://www.sciencemag.org/cgi/doi/10.1126/science.277.5331.1453. 7

M Böhm, I Wieland, K Schütze, and H Rübben. Microbeam MOMeNT: non-contact laser microdissection of membrane-mounted native tissue. *The American journal of pathology*, 151(1):63–67, 1997. 22

D W Burt. Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, 96(1-4):97–112, 2002. ISSN 14248581. URL http://www.ncbi.nlm.nih.gov/pubmed/12438785. 25

Mark J Chaisson and Pavel A Pevzner. Short read fragment assembly of bacterial genomes. *Genome Research*, 18(2):324–330, 2008. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2203630&tool=pmcentrez&rendertype=abstract. 36

L A Chelysheva, I V Solove, A V Rodionov, A F Iakovlev, and E R Gaginskaia. The lampbrush chromosomes of the chicken. cytological maps of the macrobivalents. *Tsitologiia*, 32(4):303–316, 1990. 25

M Chevalier. Donnes nouvelles sur le caryotype du lzard vivipare (Reptile, Lacer-tilien). Existe-t-il une htrogametie femelle de type Z1Z2W? . *C R Acad Sci Paris*, 268:20982100, 1969. 27

J A Coyne. Genetics and speciation. *Nature*, 355(6360):511–515, 1992. ISSN 14764687. doi: 10.1038/355511a0. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=1741030. 28

J A Coyne and H A Orr. Two rules of speciation. In D Otte and J A Endler, editors, *Speciation and its Consequences*, chapter 8, pages 180–207. Sinauer Associates, 1989. 28

Stephanie Curran and Graeme I Murray. An introduction to laser-based tissue microdissection techniques. *Methods In Molecular Biology Clifton Nj*, 293(5):3–8, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16028406. 22, 23

R Dallai and C Baroni Urbani. Fine resolution of the karyogram of Lacerta sicula campestris (DE BETTA). *Caryologia*, 20(4):347–353, 1967. 27

I S Darevsky. Natural Parthenogenesis in a Polymorphic Group of Caucasian Rock Lizards Related to Lacerta saxicola Eversmann. *JOhioHerpetolSoc*, 5(4):115–152, 1966. ISSN 04739868. URL http://www.jstor.org/stable/1562588. 27

Frank B Dean, John R Nelson, Theresa L Giesler, and Roger S Lasken. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. *Genome Research*, 11(6):1095–1099, 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11381035. 23

Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, 2008. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2532726&tool=pmcentrez&rendertype=abstract. 4

Devin Dressman, Hai Yan, Giovanni Traverso, Kenneth W Kinzler, and Bert Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America*, 100 (15):8817–8822, 2003. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=166396&tool=pmcentrez&rendertype=abstract. 2

M C Edwards and R A Gibbs. Multiplex PCR: advantages, development, and applications. *Pcr Methods And Applications*, 3(4):S65–S75, 1994. URL http://www.ncbi.nlm.nih.gov/pubmed/8173510. 9

Hans Ellegren. Genomic evidence for a large-Z effect. *Proceedings of the Royal Society B Biological Sciences*, 276(1655):361–366, 2009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2674357&tool=pmcentrez&rendertype=abstract. 28, 29

M R Emmert-Buck, R F Bonner, P D Smith, R F Chuaqui, Z Zhuang, S R Goldstein, R A Weiss, and L A Liotta. Laser capture microdissection. *Science*, 274(5289):998–1001, 1996. URL http://www.sciencemag.org/content/274/5289/998.abstract. 22

Virginia Espina, Julia D Wulfkuhle, Valerie S Calvert, Amy VanMeter, Weidong Zhou, George Coukos, David H Geho, Emanuel F Petricoin, and Lance A Liotta. Laser-capture microdissection. *Nature Protocols*, 1(2):586–603, 2006. ISSN 17502799. doi: 10.1038/nprot.2006.85. URL http://www.ncbi.nlm.nih.gov/pubmed/17406286. 22

Virginia Espina, Michael Heiby, Mariaelena Pierobon, and Lance A Liotta. Laser capture microdissection technology. *Expert Review of Molecular Diagnostics*, 7(5):647–657, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17892370. 22

Paul D Etter, Jessica L Preston, Susan Bassham, William A Cresko, and Eric A Johnson. Local De Novo Assembly of RAD Paired-End Contigs Using Short Sequencing Reads. *PLoS ONE*, 6(4):10, 2011. URL http://dx.plos.org/10.1371/journal.pone.0018561. 9

T Ezaz, S D Sarre, D Meally, J A Marshall Graves, and A Georges. Sex chromosome evolution in lizards: independent origins and rapid transitions. *Cytogenetic and Genome Research*, 127(2-4): 249–260, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/20332599. 26

Tariq Ezaz, Alexander E Quinn, Ikuo Miura, Stephen D Sarre, Arthur Georges, Jennifer A Marshall Graves, and Comparative Genomics Group. The dragon lizard Pogona vitticeps has ZZ/ZW micro-sex chromosomes. *Chromosome Research*, 13(8):763–76, 2005. ISSN 09673849. doi: 10.1007/s10577-005-1010-9. URL http://www.ncbi.nlm.nih.gov/pubmed/16331408. 32

Valérie Fillon. The chicken as a model to study microchromosomes in birds: a review. *Genetics selection evolution GSE*, 30(3):209–219, 1998. URL http://www.gsejournal.org/content/30/3/209. 25

R A Fisher. *The Genetical Theory of Natural Selection*, volume 21 of *Clarendon Press*. Clarendon Press, 1930. ISBN 0198504403. doi: 10.1038/158453a0. URL http://www.archive.org/details/geneticaltheoryo031631mbp. 28

Paul Flicek and Ewan Birney. Sense from sequence reads : methods for alignment and assembly. *Online*, 6(11), 2010. doi: 10.1038/NmEtH.1376. 3

Simon Fredriksson, Johan Banér, Fredrik Dahl, Angela Chu, Hanlee Ji, Katrina Welch, and Ronald W Davis. Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Research*, 35(7):e47, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17317684. 9

Ken Garber. Fixing the front end. *Nature Biotechnology*, 26(10):1101–4, 2008. ISSN 15461696. doi: 10.1038/nbt1008-1101. URL http://www.ncbi.nlm.nih.gov/pubmed/18846081. 10

Andreas Gnirke, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M LeProust, William Brockman, Timothy Fennell, Georgia Giannoukos, Sheila Fisher, Carsten Russ, Stacey Gabriel, David B Jaffe, Eric S Lander, and Chad Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2):182–189, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19182786. 10

D K Griffin, F Haberman, J Masabanda, P O'Brien, M Bagga, A Sazanov, J Smith, D W Burt, M Ferguson-Smith, and J Wienberg. Micro- and macrochromosome paints generated by flow cytometry and microdissection: tools for mapping the chicken genome. *Cytogenetics and Cell Genetics*, 87(3-4):278–281, 1999. URL http://www.ncbi.nlm.nih.gov/pubmed/10702695. 8

J B S Haldane. Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, 12(2):101–109, 1922. URL http://www.springerlink.com/index/L80117J46621W442.pdf. 28

Paul Hardenbol, Johan Banér, Maneesh Jain, Mats Nilsson, Eugeni A Namsaraev, George A Karlin-Neumann, Hossein Fakhrai-Rad, Mostafa Ronaghi, Thomas D Willis, Ulf Landegren, and Ronald W Davis. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology*, 21(6):673–678, 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12730666. 9

Olivier Harismendy and Kelly Frazer. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *BioTechniques*, 46(3):229–31, March 2009. ISSN 0736-6205. doi: 10.2144/000113082. URL http://www.ncbi.nlm.nih.gov/pubmed/19317667. 4

S C Harvey, J Masabanda, L A P Carrasco, N R Bromage, D J Penman, and D K Griffin. Molecular-cytogenetic analysis reveals sequence differences between the sex chromosomes of Oreochromis

niloticus: evidence for an early stage of sex-chromosome differentiation. *Cytogenetic and Genome Research*, 97(1-2):76–80, 2002. URL http://kar.kent.ac.uk/12471/. 8

I M Hastings. Manifestations of sexual selection may depend on the genetic basis of sex determination. *Proceedings of the Royal Society of London Series B Biological Sciences*, 258(1351):83–87, 1994. ISSN 09628452. URL http://www.jstor.org/stable/49978. 29

Uffe Hellsten, Richard M Harland, Michael J Gilchrist, David Hendrix, Jerzy Jurka, Vladimir Kapitonov, Ivan Ovcharenko, Nicholas H Putnam, Shengqiang Shu, Leila Taher, Ira L Blitz, Bruce Blumberg, Darwin S Dichmann, Inna Dubchak, Enrique Amaya, John C Detter, Russell Fletcher, Daniela S Gerhard, David Goodstein, Tina Graves, Igor V Grigoriev, Jane Grimwood, Takeshi Kawashima, Erika Lindquist, Susan M Lucas, Paul E Mead, Therese Mitros, Hajime Ogino, Yuko Ohta, Alexander V Poliakov, Nicolas Pollet, Jacques Robert, Asaf Salamov, Amy K Sater, Jeremy Schmutz, Astrid Terry, Peter D Vize, Wesley C Warren, Dan Wells, Andrea Wills, Richard K Wilson, Lyle B Zimmerman, Aaron M Zorn, Robert Grainger, Timothy Grammer, Mustafa K Khokha, Paul M Richardson, and Daniel S Rokhsar. The genome of the Western clawed frog Xenopus tropicalis. *Science*, 328(5978):633–636, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20431018. 25

Frederico Henning, Vladimir Trifonov, and Lurdes Foresti De Almeida-toledo. Use of chromosome microdissection in fish molecular cytogenetics. *Genetics and Molecular Biology*, 1:279–283, 2008. 8, 33

Holly H Hogrefe and Michael C Borns. Long-range PCR with a DNA polymerase fusion. *Methods In Molecular Biology Clifton Nj*, 687:17–23, 2011. 9

Austin L Hughes and Helen Piontkivska. DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size. *BMC Evolutionary Biology*, 5(1):12, 2005. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=548695&tool=pmcentrez&rendertype=abstract. 5

Susan M Huse, Julie A Huber, Hilary G Morrison, Mitchell L Sogin, and David Mark Welch. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7):R143, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17659080. 4

Illumina. DNA Sequencing with Solexa Technology. *Dna Sequenc*, 1(Figure 2):92, 2007. 4

Vikram K Iyengar, H Kern Reeve, and Thomas Eisner. Paternal inheritance of a female moth's mating preference. *Nature*, 419(6909):830–832, 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12397356. 29

Alla Katsnelson, Nik Spencer, Nick Loman, and James Hadfield. Human genome: Genomes by the thousand. *Nature*, 467(7319):1026–1027, 2010. ISSN 14764687. doi: 10.1038/4671026a. URL http://www.ncbi.nlm.nih.gov/pubmed/20981067. 4

Martin Krzywinski, Jacqueline Schein, Inan Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2752132&tool=pmcentrez&rendertype=abstract. 54

L. A. Kupryanova. [Karyological analysis of lizards of the subgenus Archaeolacerta]. *Tsitologiia*, 11: 803–814, Jul 1969. 27

L A Kupryanova and V N Arronet. Description of the karyotype of the lizard Eremias velox. *Tsitologiia*, 11(8):1057–1060, 1969. 27

E S Lander and M S Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–9, April 1988. ISSN 0888-7543. URL `http://www.ncbi.nlm.nih.gov/pubmed/3294162`. 6

L C Lawrie, S Curran, H L McLeod, J E Fothergill, and G I Murray. Application of laser capture microdissection and proteomics in colon cancer. *Molecular pathology MP*, 54(4):253–258, 2001. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1187077&tool=pmcentrez&rendertype=abstract`. 22

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract`. 53

A Lörincz. Hybrid capture., 1998. 9

H J Lüdecke, G Senger, U Claussen, and B Horsthemke. Cloning defined regions of the human genome by microdissection of banded chromosomes and enzymatic amplification. *Nature*, 338 (6213):348–350, 1989. URL `http://www.ncbi.nlm.nih.gov/pubmed/2784197`. 22, 23

Lira Mamanova, Alison J Coffey, Carol E Scott, Iwanka Kozarewa, Emily H Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J Turner. Target-enrichment strategies for next-generation sequencing. *Nature methods*, 7(2):111–8, February 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1419. URL `http://www.ncbi.nlm.nih.gov/pubmed/20111037`. 10

Judith E Mank, David W Hall, Mark Kirkpatrick, and John C Avise. Sex chromosomes and male ornaments: a comparative evaluation in ray-finned fishes. *Proceedings of the Royal Society B Biological Sciences*, 273(1583):233–236, 2006. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560031&tool=pmcentrez&rendertype=abstract`. 29

M Margulies, M Egholm, W E Altman, S Attiya, J S Bader, L A Bemben, J Berka, M S Braverman, Y J Chen, Z Chen, S B Dewell, L Du, J M Fierro, X V Gomes, B C Godwin, W He, S Helgesen, C H Ho, C H Ho, G P Irzyk, S C Jando, M L Alenquer, T P Jarvie, K B Jirage, J B Kim, J R Knight, J R Lanza, J H Leamon, S M Lefkowitz, M Lei, J Li, K L Lohman, H Lu, V B Makhijani, K E McDade, M P McKenna, E W Myers, E Nickerson, J R Nobile, R Plant, B P Puc, M T Ronan, G T Roth, G J Sarkis, J F Simons, J W Simpson, M Srinivasan, K R Tartaro, A Tomasz, K A Vogt, G A Volkmer, S H Wang, Y Wang, M P Weiner, P Yu, R F Begley, and J M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/16056220`. 2

R Matthey. Chromosomes de Reptiles Sauriens, Ophidiens, Chélonien. L'évolution de la formule chromosomiale chez les Sauriens. *Revue Suisse De Zoologie*, 40(3):117–185, 1931. ISSN 0035418X. 27

R. Matthey. La loi de Robertson et la formule chromosomiale chez deux lacertiens: Lacerta oceltata Darts.,Psammodromus hispanicus Fitz. *Cytologia*, 10:32–39, December 1939a. 27

R. Matthey. L 'evolution de la formule chromosomiale chez les vertebres. *Experientia*, 1:50–56, 78–86, December 1939b. 27

Paul Medvedev, Monica Stanciu, and Michael Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11 Suppl):S13–S20, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19844226. 5

P S Meltzer, X Y Guan, A Burgess, and J M Trent. Rapid generation of region specific probes by chromosome microdissection and their application. *Nature Genetics*, 1(1):24–28, 1992. 22

Florian Mertes, Abdou Elsharawy, Sascha Sauer, Joop M L M Van Helvoort, P J Van Der Zaag, Andre Franke, Mats Nilsson, Hans Lehrach, and Anthony J Brookes. Targeted enrichment of genomic dna regions for next-generation sequencing. *Briefings in functional genomics*, 10(6):374–86, 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/22121152. 7

P Métézeau, A Schmitz, and G Frelat. Analysis and sorting of chromosomes by flow cytometry: new trends. *Biology of the cell under the auspices of the European Cell Biology Organization*, 78(1-2): 31–39, 1993. 22

Michael L Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1): 31–46, January 2010. ISSN 1471-0064. doi: 10.1038/nrg2626. URL http://www.ncbi.nlm.nih.gov/pubmed/19997069. 3

Linda Strömqvist Meuzelaar, Owen Lancaster, J Paul Pasche, Guido Kopal, and Anthony J Brookes. MegaPlex PCR: a strategy for multiplex amplification. *Nature Methods*, 4(10):835–837, 2007. URL http://hdl.handle.net/2381/3833. 9

Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–27, June 2010. ISSN 1089-8646. doi: 10.1016/j.ygeno.2010.03.001. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2874646&tool=pmcentrez&rendertype=abstract. 11, 17, 36, 45

Michael R Miller, Joseph P Dunham, Angel Amores, William A Cresko, and Eric A Johnson. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2):240–8, 2007. ISSN 10889051. doi: 10.1101/gr.5681207. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1781356&tool=pmcentrez&rendertype=abstract. 9

Patrice Milos. Helicos BioSciences. *Pharmacogenomics*, 9(4):477–480, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18384261. 2

R D Mitra and G M Church. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research*, 27(24):e34, 1999. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148757&tool=pmcentrez&rendertype=abstract. 2

E W Myers. Toward simplifying and accurately formulating fragment assembly. *Journal of computational biology a journal of computational molecular cell biology*, 2(2):275–290, 1995. URL http://www.ncbi.nlm.nih.gov/pubmed/7497129. 18

Atsushi Nakabachi, Atsushi Yamashita, Hidehiro Toh, Hajime Ishikawa, Helen E Dunbar, Nancy A Moran, and Masahira Hattori. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science*, 314(5797):267, 2006. ISSN 10959203. doi: 10.1126/science.1134196. URL http://www.sciencemag.org/cgi/content/abstract/314/5797/267. 3

M Nilsson, H Malmgren, M Samiotaki, M Kwiatkowski, B P Chowdhary, and U Landegren. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science*, 265(5181):2085–2088, 1994. URL http://www.ncbi.nlm.nih.gov/pubmed/7522346. 10

YK Oh. A karyotype study in chiroptera (bats). *Yonsei Med*, 16(2):46–53, 1975. URL http://www.ncbi.nlm.nih.gov/pubmed/1232711. 25

V. F. Orlova and N. F. Orlov. [Chromosome complements and some questions of systematics of lizards of the genus Lacerta. *Zool. Zh.*, 48:1056–1060, Jul 1969. 27

Kit Part, Ship Kit Part, Additional Equipment, and Reagents Required. GS FLX Titanium Rapid Library Preparation Kit. *Change*, 2010. 34

Mihaela Pavlicev and Werner Mayer. Fast radiation of the subfamily Lacertinae (Reptilia: Lacertidae): history or methodical artefact? *Molecular Phylogenetics and Evolution*, 52(3):727–734, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19427911. 26

Jaume Pellicer, Michael F Fay, Royal Botanic Gardens, and Richmond Surrey Tw. The largest eukaryotic genome of them all ? *Society*, 164(1):10–15, 2010. ISSN 00244074. doi: 10.1111/j.1095-8339.2010.01072.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8339.2010.01072.x/full. 3

Catarina Pinho, Sara Rocha, Bruno Carvalho, Susana Lopes, Sofia Mouro, Marcelo Vallinoto, Tuliana Brunes, Clio Haddad, Helena Gonalves, Fernando Sequeira, and Nuno Ferrand. New primers for the amplification and sequencing of nuclear loci in a taxonomically wide set of reptiles and amphibians. *Conservation Genetics Resources*, 2:181–185, 2010. ISSN 1877-7252. URL http://dx.doi.org/10.1007/s12686-009-9126-4. 10.1007/s12686-009-9126-4. 68

Martina Pokorná, Marie Rábová, Petr Ráb, Malcolm a Ferguson-Smith, Willem Rens, and Lukáš Kratochvíl. Differentiation of sex chromosomes and karyotypic evolution in the eye-lid geckos (Squamata: Gekkota: Eublepharidae), a group with different modes of sex determination. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 18(7):809–20, November 2010. ISSN 1573-6849. doi: 10.1007/s10577-010-9154-7. URL http://www.ncbi.nlm.nih.gov/pubmed/20811940. 33

Martin F Polz and Colleen M Cavanaugh. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, 64(10):3724–3730, 1998. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=106531&tool=pmcentrez&rendertype=abstract. 10

Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, 10(4):354–366, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19482960. 17

D P Prowell. Sex linkage and speciation in Lepidoptera. *The Canadian Entomologist*, 126:309–319, 1998. 29

Petr Ráb, Marie Rábová, Carla Sofia Pereira, Maria João Collares-Pereira, and Sárka Pelikánová. Chromosome studies of European cyprinid fishes: interspecific homology of leuciscine cytotaxonomic marker-the largest subtelocentric chromosome pair as revealed by cross-species painting. *Chromosome research an international journal on the molecular supramolecular and evolutionary aspects of chromosome biology*, 16(6):863–873, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18709543`. 33

K Reinhold. Sex linkage among genes controlling sexually selected traits. *Behavioral Ecology and Sociobiology*, 44(1):1–7, 1998. ISSN 14346621. URL `http://www.springerlink.com/index/XVV905URA9539FD1.pdf`. 29

A. V. Rodionov. [Micro vs. macro: structural-functional organization of avian micro- and macrochromosomes]. *Genetika*, 32:597–608, May 1996. 26

A V Rodionov, L A Chelsheva, I V Solove, and Iu A Miakoshina. Chiasma distribution in the lampbrush chromosomes of the chicken gallus gallus domesticus: hot spots of recombination and their possible role in proper dysjunction of homologous chromosomes at the first meiotic division. *Genetika*, 28(7):151–160, 1992. 25

Adrianna S Rodriguez, Benjamin H Espina, Virginia Espina, and Lance A Liotta. Automated laser capture microdissection for tissue proteomics. *Methods In Molecular Biology Clifton Nj*, 441:71–90, 2008. 22

S Rozen and H J Skaletsky. Primer3. *Bioinformatics Methods and Protocols Methods in Molecular Biology*, 3(0.9):1–41, 1998. URL `http://www-genome.wi.mit.edu/genome_software/other/primer3.html`. 67

RK Saiki, DH Gelfand, S Stoffel, and ST Scharf. Primer-directed enzymatic amplification of DNA. *Science*, 239:487–491, 1988. URL `http://www.etseq.urv.es/doctorat/index/web_nanobiotech/handouts/Lecture_7_Handout_7_PCR_publication.pdf`. 9

J Sambrook, E F Fritsch, and T Maniatis. *Molecular Cloning: A Laboratory Manual. Cold Spring Harbor laboratory press*. Cold Spring Harbor, 1989. 68

F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, C A Fiddes, C A Hutchison, P M Slocombe, and M Smith. 1977 Nature Publishing Group. *Group*, 265(5596):687–95, 1977. ISSN 00280836. doi: 10.1038/266309a0. URL `http://www.ncbi.nlm.nih.gov/pubmed/870828`. 2

F Scalenghe, E Turco, J E Edstrom, V Pirrotta, and M Melli. Microdissection and cloning of DNA from a specific region of Drosophila melanogaster polytene chromosomes. *Chromosoma*, 82(2): 205–216, 1981. URL `http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=0006452995`. 22

Stephen F Schaffner. The X chromosome in population genetics. *Nature reviews. Genetics*, 5(1): 43–51, January 2004. ISSN 1471-0056. doi: 10.1038/nrg1247. URL `http://www.ncbi.nlm.nih.gov/pubmed/14708015`. 27

Michael C Schatz, Arthur L Delcher, and Steven L Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Research*, 20(9):1165–1173, 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20508146`. 16, 17

K Schütze and G Lahr. Identification of expressed genes by laser-mediated manipulation of single cells. *Nature Biotechnology*, 16(8):737–742, 1998. URL http://dx.doi.org/10.1038/nbt0898-737. 22

G Senger, H J Lüdecke, B Horsthemke, and U Claussen. Microdissection of banded human chromosomes. *Human Genetics*, 84(6):507–511, 1990. 22

N L Simone, R F Bonner, J W Gillespie, M R Emmert-Buck, and L A Liotta. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends in Genetics*, 14 (7):272–276, 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/9676529. 22

J Smith, C K Bruley, I R Paton, I Dunn, C T Jones, D Windsor, D R Morrice, A S Law, J Masabanda, A Sazanov, D Waddington, R Fries, and D W Burt. Differences in gene density on chicken macrochromosomes and microchromosomes. *Animal Genetics*, 31(2):96–103, 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10782207. 25

Roger Staden. Nucleic Acids Research Automation of the computer handling of gel reading data produced by the shotgun method of Nucleic Acids Research. *Nucleic Acids Research*, 1(15):4731–4751, 1982. 6

Jessica Stapley, Julia Reger, Philine G D Feulner, Carole Smadja, Juan Galindo, Robert Ekblom, Clair Bennison, Alexander D Ball, Andrew P Beckerman, and Jon Slate. Adaptation genomics: the next generation. *Trends in ecology & evolution*, 25(12):705–12, December 2010. ISSN 0169-5347. doi: 10.1016/j.tree.2010.09.002. URL http://www.ncbi.nlm.nih.gov/pubmed/20952088. 9

Daniel Summerer, Haiguo Wu, Bettina Haase, Yang Cheng, Nadine Schracke, Cord F Stähler, Mark S Chee, Peer F Stähler, and Markus Beier. Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Research*, 19(9):1616–21, 2009. ISSN 15495469. doi: 10.1101/gr.091942.109. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2752126&tool=pmcentrez&rendertype=abstract. 10

Jamie K Teer, Lori L Bonnycastle, Peter S Chines, Nancy F Hansen, Natsuyo Aoyama, Amy J Swift, Hatice Ozel Abaan, Thomas J Albert, Elliott H Margulies, Eric D Green, Francis S Collins, James C Mullikin, and Leslie G Biesecker. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome research*, 20(10):1420–31, October 2010. ISSN 1549-5469. doi: 10.1101/gr.106716.110. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2945191&tool=pmcentrez&rendertype=abstract. 7, 10

H Telenius, N P Carter, C E Bebb, M Nordenskjöld, B A Ponder, and A Tunnacliffe. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics*, 13(3):718–725, 1992. URL http://www.ncbi.nlm.nih.gov/pubmed/1639399. 8, 24

Michael Turelli and Leonie C Moyle. Asymmetric postmating isolation: Darwin's corollary to Haldane's rule. *Genetics*, 176(2):1059–1088, 2007. ISSN 00166731. URL http://www.genetics.org/cgi/content/abstract/176/2/1059. 28

Emily H Turner, Sarah B Ng, Deborah a Nickerson, and Jay Shendure. Methods for genomic partitioning. *Annual review of genomics and human genetics*, 10:263–84, January 2009. ISSN 1545-293X. doi: 10.1146/annurev-genom-082908-150112. URL http://www.ncbi.nlm.nih.gov/pubmed/19630561. 10

Alex Van Belkum, Stewart Scherer, Loek Van Alphen, and Henri Verbrugh. Short-Sequence DNA Repeats in Prokaryotic Genomes. *Microbiology and Molecular Biology Reviews*, 62(2):275–293, 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/10673001. 5

J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15001713. 17

Beatriz Vicoso and Brian Charlesworth. Evolution on the X chromosome: unusual patterns and processes. *Nature reviews. Genetics*, 7(8):645–53, August 2006. ISSN 1471-0056. doi: 10.1038/nrg1914. URL http://www.ncbi.nlm.nih.gov/pubmed/16847464. 27

J N Volff and M Schartl. Variability of genetic sex determination in poeciliid fishes. *Genetica*, 111 (1-3):101–110, 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11841158. 29

P M Warnecke, C Stirzaker, J R Melki, D S Millar, C L Paul, and S J Clark. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Research*, 25(21):4422–4426, 1997. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=147052&tool=pmcentrez&rendertype=abstract. 9

Nava Whiteford, Niall Haslam, Gerald Weber, Adam Prügel-Bennett, Jonathan W Essex, Peter L Roach, Mark Bradley, and Cameron Neylon. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, 33(19):e171, 2005. URL http://eprints.ecs.soton.ac.uk/11605/1/e171. 4, 7

J Yu, S Tong, T Yang-Feng, and F T Kao. Construction and characterization of a region-specific microdissection library from human chromosome 2q35-q37. *Genomics*, 14(3):769–774, 1992. 22

A Zahavi. Mate Selection - Selection for a Handicap. *Journal of Theoretical Biology*, 53(1):205–214, 1975. ISSN 00225193. 28

L Zhang, X Cui, K Schmitt, R Hubert, W Navidi, and N Arnheim. Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):5847–5851, 1992. URL http://www.ncbi.nlm.nih.gov/pubmed/21738716. 23

Ruo-Nan Zhou and Zan-Min Hu. The development of chromosome microdissection and microcloning technique and its applications in genomic research. *Current genomics*, 8(1):67–72, March 2007. ISSN 1389-2029. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2474687&tool=pmcentrez&rendertype=abstract. 8, 21, 23

R Zimmer, W A King, and A M Verrinder Gibbins. Generation of chicken Z-chromosome painting probes by microdissection for screening large-insert genomic libraries. *Cytogenetics and Cell Genetics*, 78(2):124–130, 1997. 33

Supplementary images

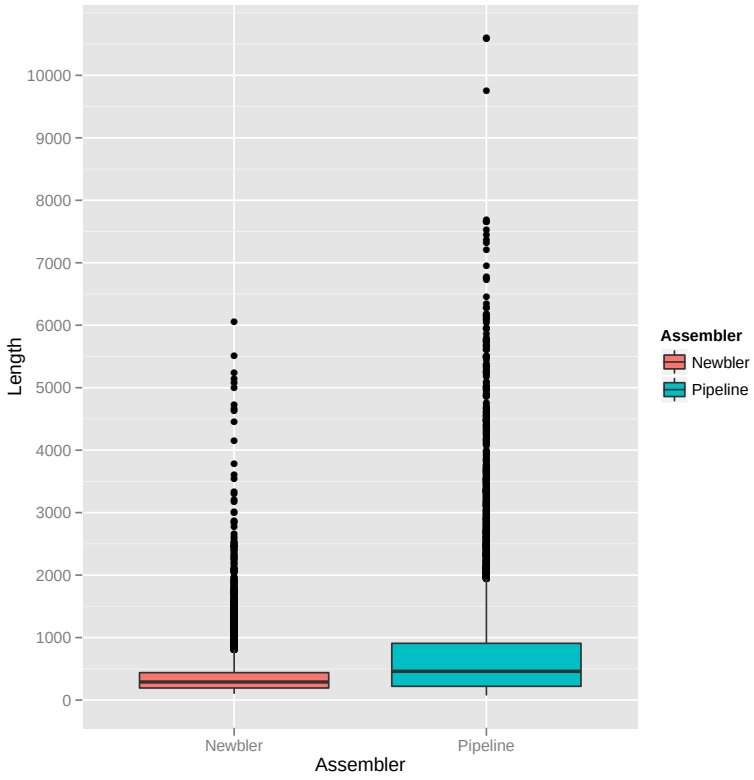


Figure 8.1: Boxplot of contig length distribution for both assemblies.

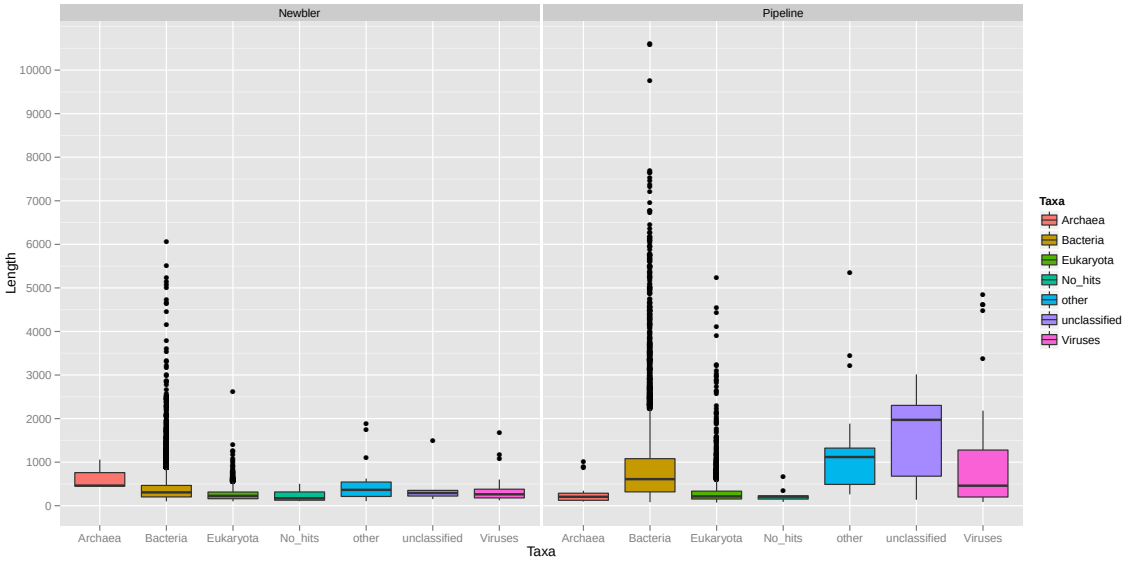


Figure 8.2: Boxplot with contig length distribution by taxa obtained from the contigs top BLAST hit.

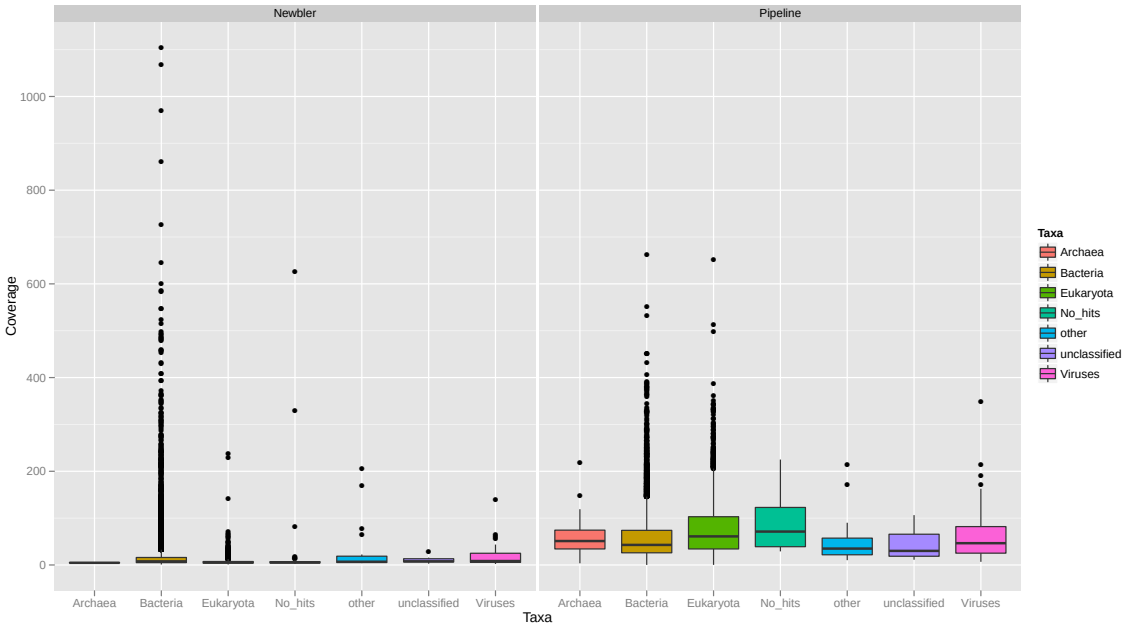


Figure 8.3: Boxplot for depth of coverage distribution by taxa for Newbler (on the left), and Mathematica Pipeline (on the right).

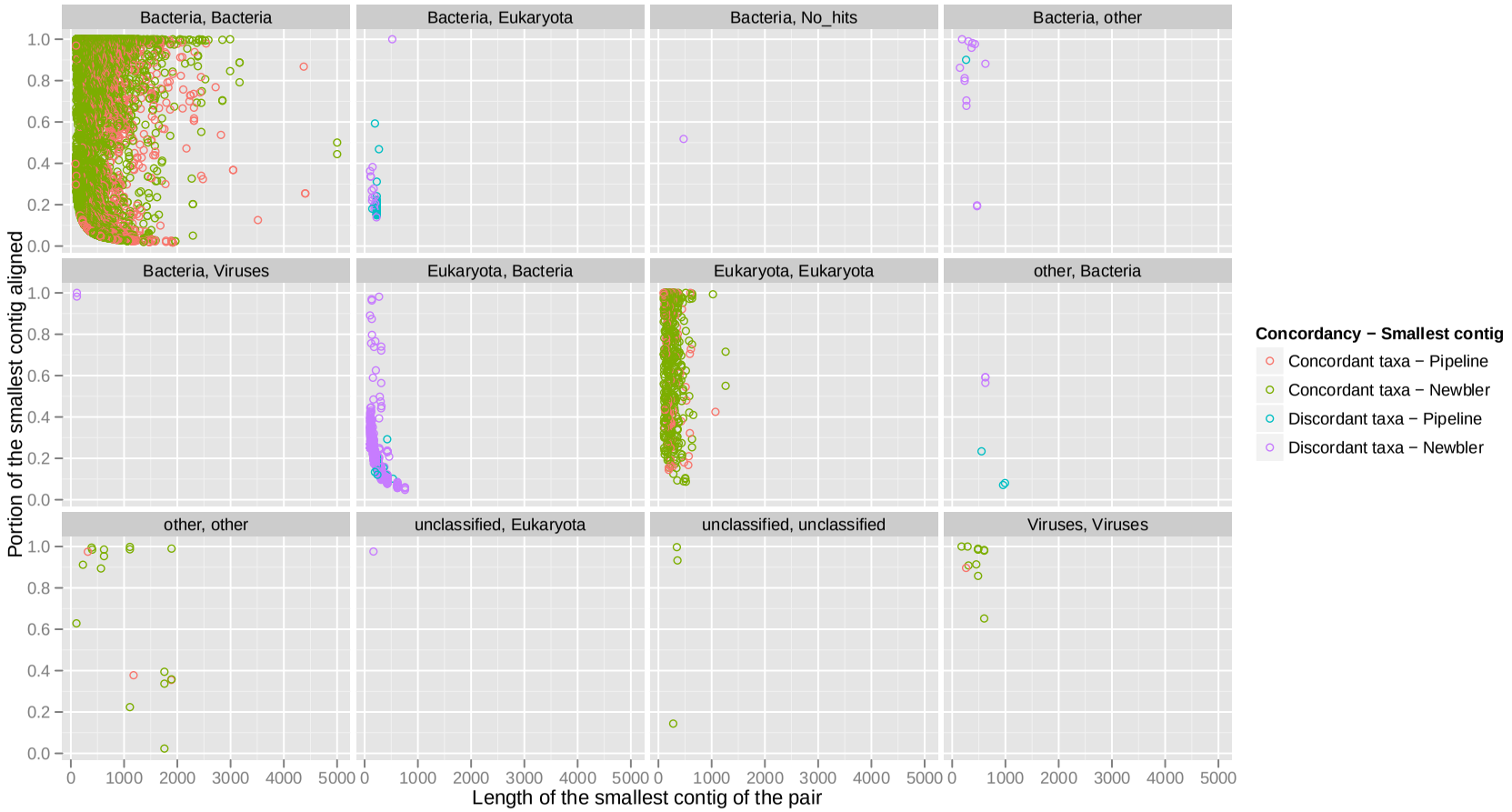


Figure 8.4: **Scatter plots displaying all possible combinations of BLAST results between the two best matching contigs from each assembly, which had aligned over at least 50% of their width to its assigned taxon.** Plots with pink and green dots display those contigs for which the taxa is concordant between the two assemblies. Plots with blue and purple dots represent those contigs for which the taxa is discordant between the two assemblies. Each dot represents the length (x axis) of the smallest of the two contigs from both assemblies, and the portion of the smallest contig of the pair aligned between the two contigs (y axis). Pink and blue when Newbler has the smallest contigs of the pair, purple and green otherwise.

Appendix A

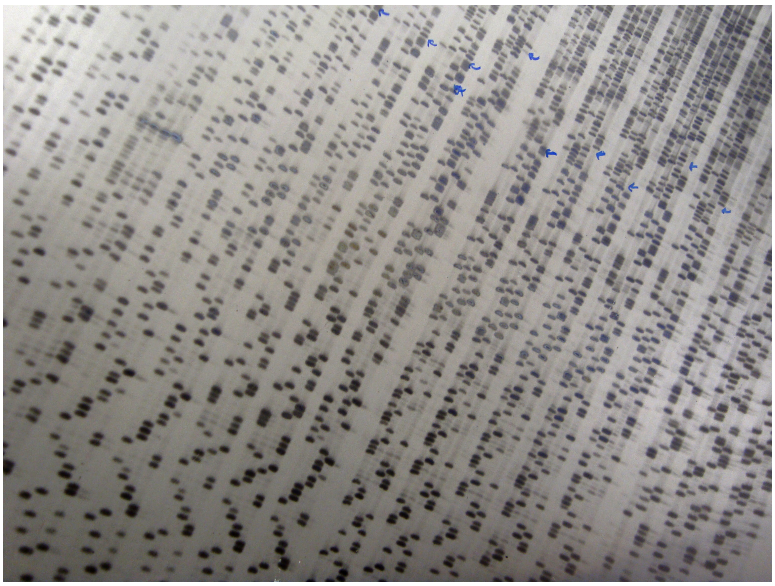


Figure A.1: **A glimpse into a not so far past: Polyacrylamide slab gel, S-35 labeled sequencing reactions.** Each block of four consecutive lanes correspond to the four letters of the genetic alphabet {A,C,G,T}.

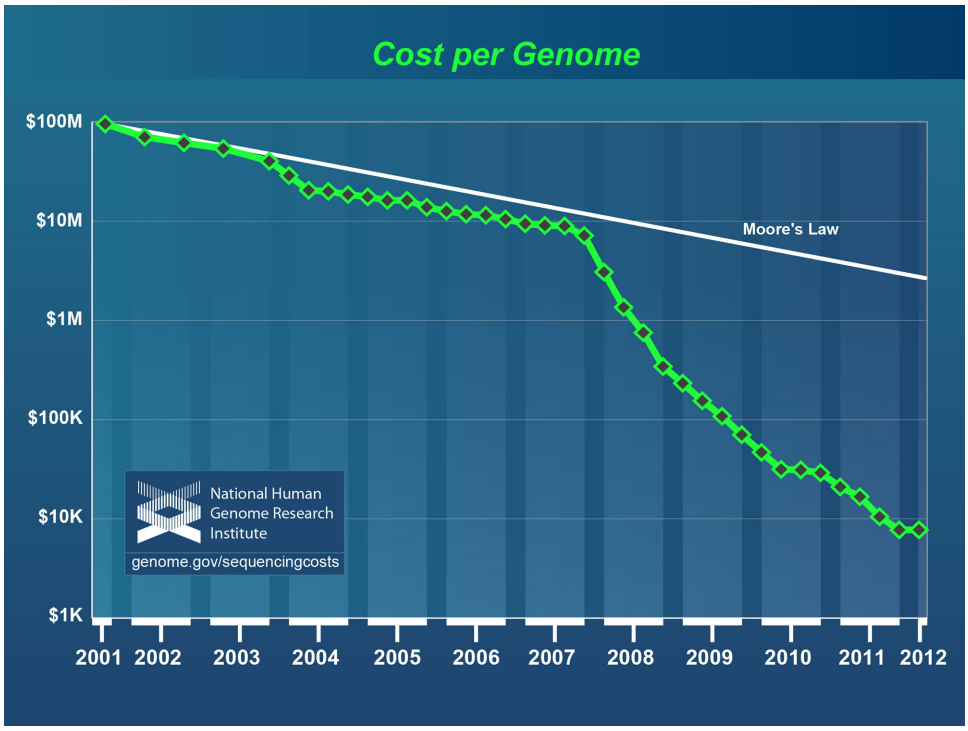


Figure A.2: **Sequencing cost evolution of a 3 billion base pairs genome compared to Moore's law from 2001 to 2012.** White line shows hypothetical data reflecting Moore's Law. This law predicts that the "compute power" doubles approximately every two years, and is often used as a measure of how well technology advancements are developing. Green line shows the reduction on the cost to sequence a genome with 3.000.000 base pairs in the past eleven years. The abrupt decline in January 2008 coincides with the time when labs started adopting the so-called next-generation sequencing platforms in detriment of the older Sanger and capillary sequencing.

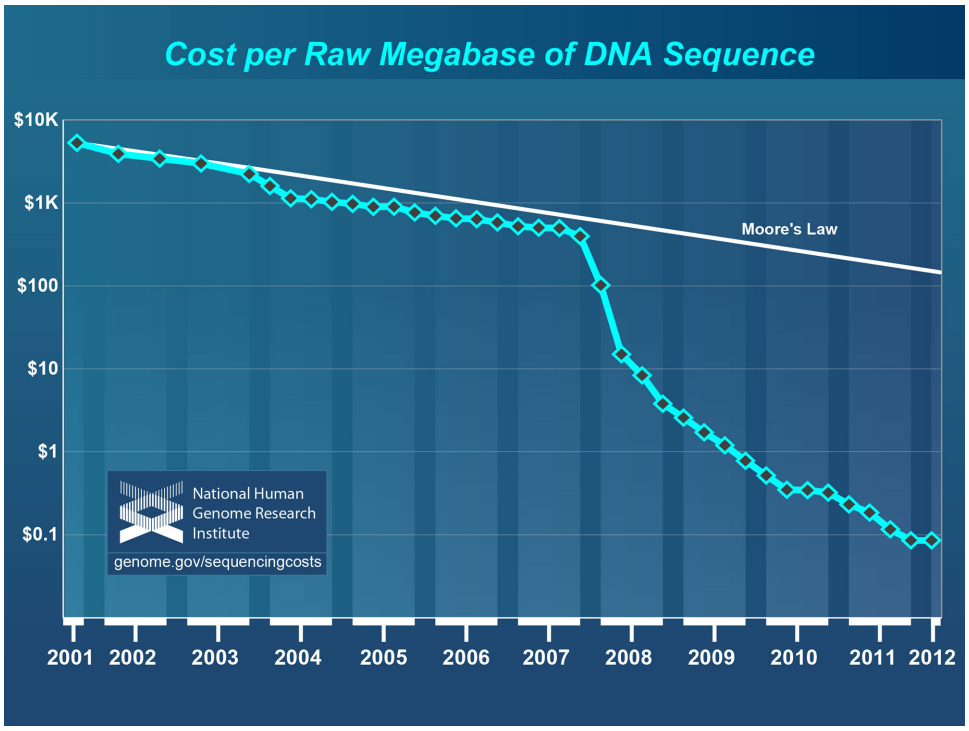


Figure A.3: **Sequencing cost per megabase compared to Moore's law from 2001 to 2012.** White line shows hypothetical data reflecting Moore's Law. This law predicts that the "compute power" doubles approximately every two years, and is often used as a measure of how well technology advancements are developing. Light blue line shows the reduction of the sequencing cost per megabase in the past eleven years. The abrupt decline in January 2008 coincides with the time when labs started adopting the so-called next-generation sequencing platforms in detriment of the older Sanger and capillary sequencing.

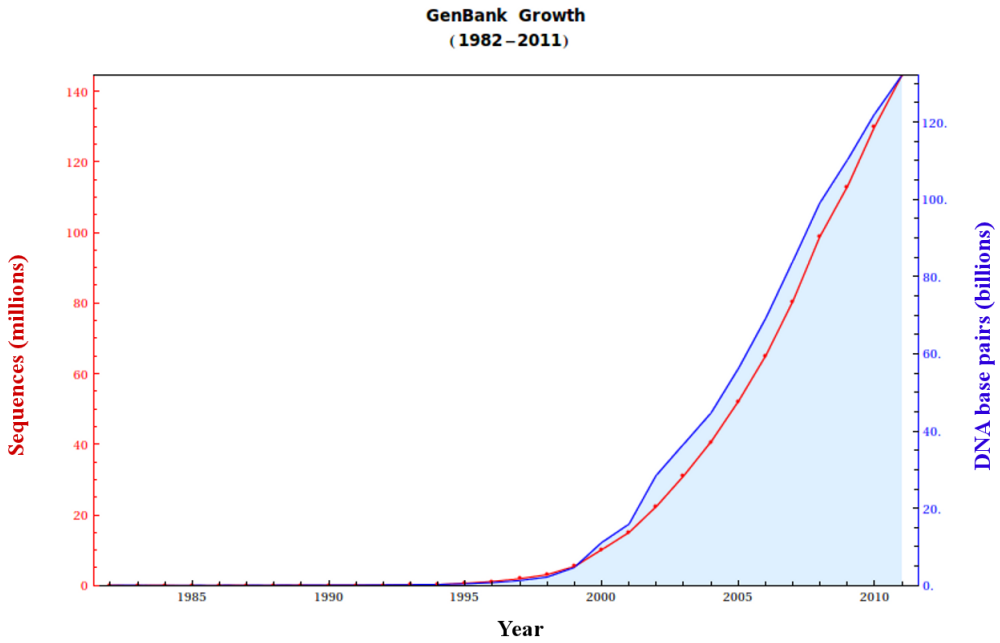


Figure A.4: **GenBank growth from December 1982 to October 2011.** The red line shows the yearly growth in millions of sequences, and the blue line shows the growth in billions of base pairs. (Data available at ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt)

Figure A.5: Poster depicting the most used next-generation sequencing methods

nature REVIEWS GENETICS

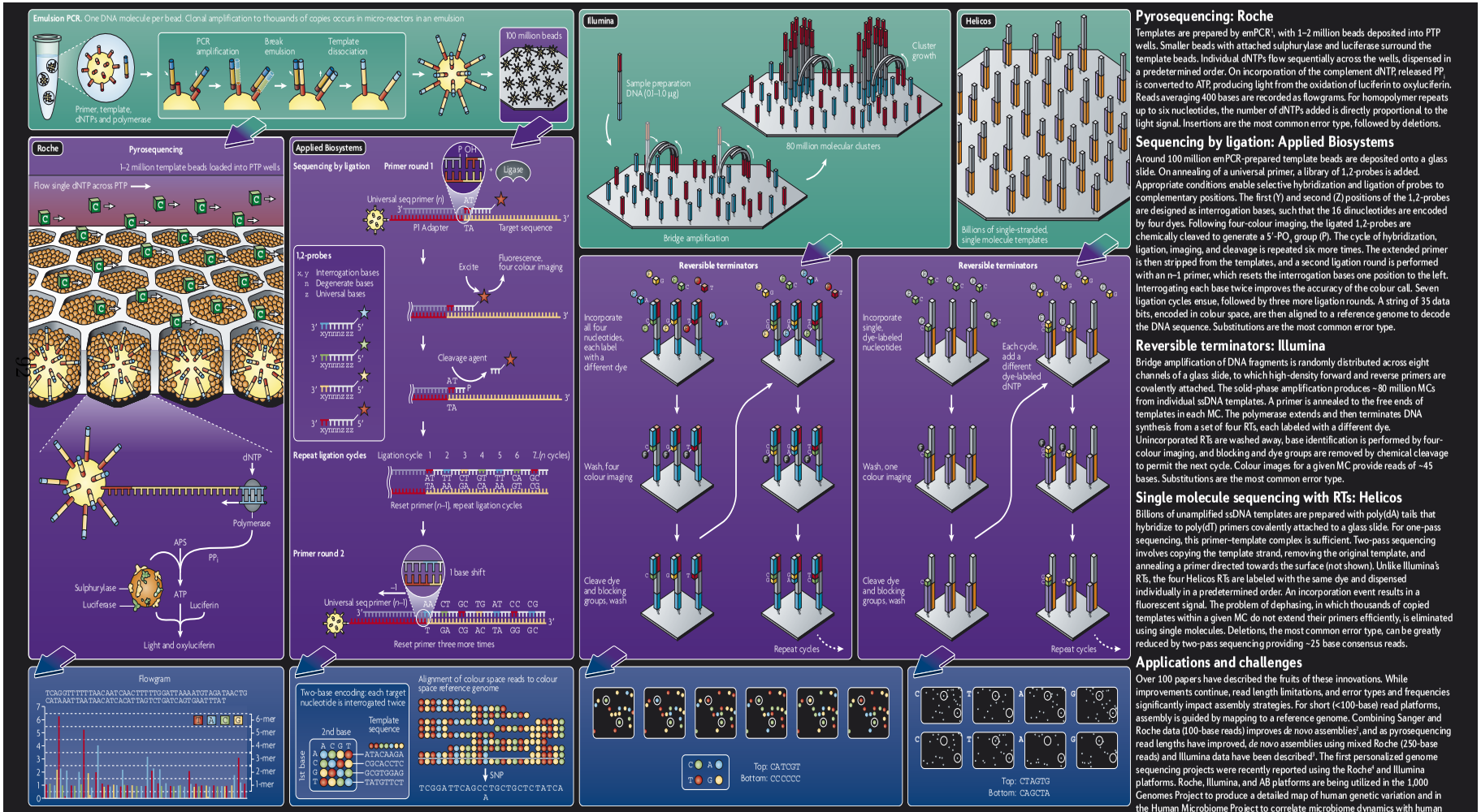
nature genetics

Sequencing technologies — the next generation

Michael L. Metzker

Never has the state of DNA sequencing technology been in greater flux than today. The steadfast approach of fluorescence-based Sanger sequencing appears to have reached its limit for technological improvements. It is being replaced by emerging technologies that promise faster and cheaper sequence information in far greater volumes than ever before. These next generation methodologies push back the limits of possibility, enabling research that would be impractical and too expensive using the Sanger paradigm. With this transition come new possibilities in the field of large-scale genomic science, coupled with new challenges in data storage and analysis. Here, the technical details of commercially available, next generation sequencing platforms are highlighted, along with their advantages and disadvantages. Gone are the days of a single platform capable of addressing the needs of most researchers. Investigators must now identify, from several DNA sequencing approaches, the one platform — or combination thereof — that best serves their application.

Applied Biosystems



Pyrosequencing: Roche

Templates are prepared by emPCR[1], with 1–2 million beads deposited into PTP wells. Smaller beads with attached sulphurylase and luciferase surround the template beads. Individual dNTPs flow sequentially across the wells, dispensed in a predetermined order. On incorporation of the complement dNTP, released PP$_i$ is converted to ATP, producing light from the oxidation of luciferin to oxyluciferin. Reads averaging 400 bases are recorded as flowgrams. For homopolymer repeats up to six nucleotides, the number of dNTPs added is directly proportional to the light signal. Insertions are the most common error type, followed by deletions.

Sequencing by ligation: Applied Biosystems

Around 100 million emPCR-prepared template beads are deposited onto a glass slide. On annealing of a universal primer, a library of 1,2-probes is added. Appropriate conditions enable selective hybridization and ligation of probes to complementary positions. The first (Y) and second (Z) positions of the 1,2-probes are designed as interrogation bases, such that the 16 dinucleotides are encoded by four dyes. Following four-colour imaging, the ligated 1,2-probes are chemically cleaved to generate a 5′-PO$_4$ group (P). The cycle of hybridization, ligation, imaging, and cleavage is repeated six more times. The extended primer is then stripped from the templates, and a second ligation round is performed with an n–1 primer, which resets the interrogation bases one position to the left. Interrogating each base twice improves the accuracy of the colour call. Seven ligation cycles ensue, followed by three more ligation rounds. A string of 35 data bits, encoded in colour space, are then aligned to a reference genome to decode the DNA sequence. Substitutions are the most common error type.

Reversible terminators: Illumina

Bridge amplification of DNA fragments is randomly distributed across eight channels of a glass slide, to which high-density forward and reverse primers are covalently attached. The solid-phase amplification produces ~80 million MCs from individual ssDNA templates. A primer is annealed to the free ends of templates in each MC. The polymerase extends and then terminates DNA synthesis from a set of four RTs, each labeled with a different dye. Unincorporated RTs are washed away, base identification is performed by four-colour imaging, and blocking and dye groups are removed by chemical cleavage to permit the next cycle. Colour images for a given MC provide reads of ~45 bases. Substitutions are the most common error type.

Single molecule sequencing with RTs: Helicos

Billions of unamplified ssDNA templates are prepared with poly(dA) tails that hybridize to poly(dT) primers covalently attached to a glass slide. For one-pass sequencing, this primer–template complex is sufficient. Two-pass sequencing involves copying the template strand, removing the original template, and annealing a primer directed towards the surface (not shown). Unlike Illumina's RTs, the four Helicos RTs are labeled with the same dye and dispensed individually in a predetermined order. An incorporation event results in a fluorescent signal. The problem of dephasing, in which thousands of copied templates within a given MC do not extend their primers efficiently, is eliminated using single molecules. Deletions, the most common error type, can be greatly reduced by two-pass sequencing providing ~25 base consensus reads.

Applications and challenges

Over 100 papers have described the fruits of these innovations. While improvements continue, read length limitations, and error types and frequencies significantly impact assembly strategies. For short (<100-base) read platforms, assembly is guided by mapping to a reference genome. Combining Sanger and Roche data (100-base reads) improves *de novo* assemblies[2], and as pyrosequencing read lengths have improved, *de novo* assemblies using mixed Roche (250-base reads) and Illumina data have been described[3]. The first personalized genome sequencing projects were recently reported using the Roche[4] and Illumina platforms. Roche, Illumina, and AB platforms are being utilized in the 1,000 Genomes Project to produce a detailed map of human genetic variation and in the Human Microbiome Project to correlate microbiome dynamics with human health. Applications are not limited to sequencing genomes. Consensus counting assays[5] have recently emerged, enabling global profiling of transcription factor binding, mRNA splicing, DNA methylation, small RNAs, chromatin structure, and DNase hypersensitivity sites. Paired-end sequencing protocols have also been developed by Roche, Illumina, and AB. These are not only important for *de novo* assemblies, but for identifying structural variation and mapping mRNA splice isoforms. Looking to the future, the development of platforms from companies such as Pacific Biosciences, Dover Systems (Polonator G.007), Visigen Biotechnologies, LaserGen, Inc., Intelligent Bio-Systems, Complete Genomics, and Oxford Nanopore Technology — as well as those described here — is expected to further improve throughput and accuracy while lowering cost.

Applied Biosystems Inc. is a global leader in the development and commercialization of instrument-based systems, consumables, software, and services for the life-science market. Beginning with the launch of the first automated DNA sequencer in 1986, Applied Biosystems revolutionized the field of genetic analysis by delivering innovative chemistries and robust instrumentation. These technologies enabled groundbreaking science including the first draft of the human genome and the genomes of >500 other organisms. With each new generation of capillary electrophoresis instrumentation, Applied Biosystems redefined the speed, cost and accuracy of sequence analysis.

The SOLiD™ System is Applied Biosystem's latest revolutionary step forward in genomic analysis technology. It provides a highly accurate, massively parallel genomic analysis platform, which supports a wide range of applications, including whole-genome sequencing, chromatin immunoprecipitation, microbial sequencing, digital karyotyping, medical sequencing, genotyping, gene expression and small RNA discovery. The SOLiD™ System's inherent scalability has demonstrated advances in throughput of >1.7GB in a single run. The flexibility of two independent flow cells allows multiple experiments to be conducted in a single run. With unparalleled throughput and >99.9% overall accuracy, the SOLiD™ System enables large-scale sequencing and tag experiments to be completed more cost effectively than previously possible.

The SOLiD™ System is supported by one of the life-science industry's most comprehensive service and support organizations of more than 2,000 dedicated field personnel worldwide, specializing in business consulting and protocol development, instrument optimization, and data and application integration. Further information about the SOLiD™ System is available at http://solid.appliedbiosystems.com.

References

[1]See Abbreviations. | [2]Goldberg, S.M.D. *et al.* A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. U S A* 103, 11240-11245 (2006). | [3]Holt, K.E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nature Genet.* 40, 987-993 (2008). | [4]Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872-876 (2008). | [5]Wold, B. & Myers, R.M. Sequence census methods for functional genomics. *Nature Methods* 5, 19-21 (2008).

Abbreviations

APS, adenosine 5′-phosphosulphate; ATP, adenosine triphosphate; dNTP, 2′-deoxyribonucleoside triphosphate; emPCR, emulsion PCR; MC, molecular cluster; PP$_i$, inorganic pyrophosphate; PTP, PicoTiterPlate; RT, reversible terminator; SNP, single nucleotide polymorphism; ssDNA, single-stranded DNA.

Acknowledgements

Michael L. Metzker is at the Human Genome Sequencing Center and Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, Texas, USA. Edited by Louisa Flintoft; designed by Patrick Morgan.

http://www.nature.com/reviews/posters/sequencing